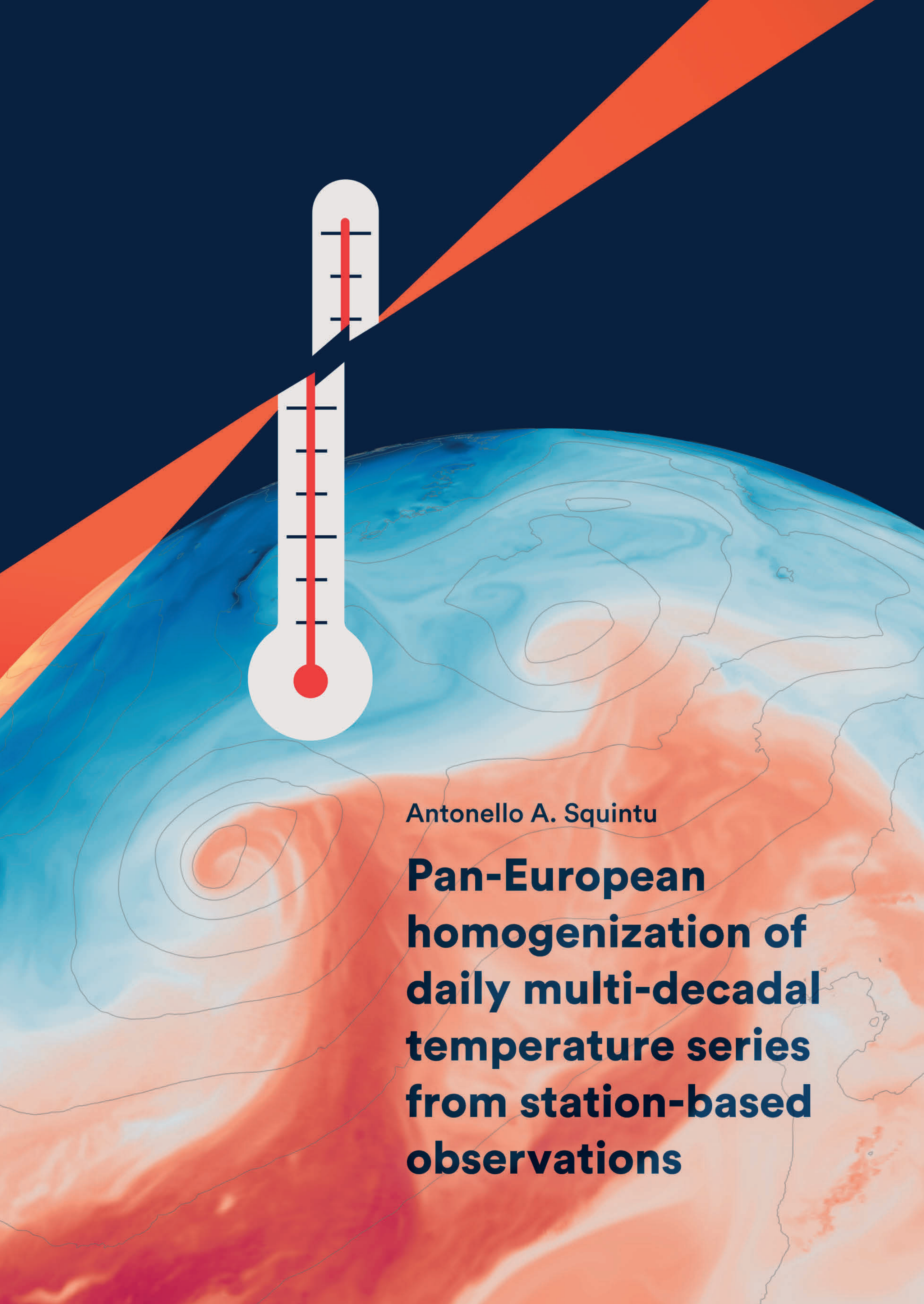


Antonello A. Squintu

Pan-European homogenization of daily multi-decadal temperature series from station-based observations

PROPOSITIONS

1. Long daily temperature series without inhomogeneities are an illusion.
   (this thesis)

2. Homogenization processes do not aim to increase or decrease warming trends in temperature series.
   (this thesis)

3. Every scientist should be aware of their own field of expertise, ignoring these limits is disrespectful.

4. Scientific conclusions cannot be taken through consensus.

5. Deadlines are one of the best incentives for human activities.

6. Accessibility to long-distance travels should be guaranteed in a sustainable society.

Propositions belonging to the thesis entitled:

Pan-European homogenization of daily multi-decadal temperature series from station-based observations

Antonello Angelo Squintu

Wageningen, 16th October 2020

Pan-European homogenization of daily multi-decadal temperature series from station-based observations

Antonello A. Squintu

Thesis committee

Promotor:
Prof.Dr Albert M. G. Klein Tank
Wageningen University & Research

Co-promotors:
Dr Gerard van der Schrier
Senior Scientist, R&DWD
Koninlijk Nederlands Meteorologisch Instituut, De Bilt

Other members:
Prof.Dr Ton Hoitink, Wageningen University & Research
Prof.Dr Enric Aguilar, Universitat Rovira i Virgili, Tarragona, Spain
Dr John H. van Boxel, Universiteit van Amsterdam
Dr Lisette (E.)J. Klok, Hogeschool van Amsterdam

This research was conducted under the auspices of the Graduate School for Socio-Economic and Natural Sciences of the Environment (SENSE).

Pan-European homogenization of daily multi-decadal temperature series from station-based observations

Antonello A. Squintu

Thesis
submitted in fulfilment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 16 October 2020
at 1:30 p.m. in the Aula.

Antonello A. Squintu
Pan-European homogenization of daily multi-decadal
temperature series from station-based observations
xvi + 162 pages.

PhD thesis, Wageningen University, Wageningen, NL (2020)
With references, with summary in English and Italian

ISBN 978-94-6395-473-0
DOI https://doi.org/10.18174/527688

Al mio

pensiero felice

Summary

The changes of European climate have serious effects on society and economy. A thorough climatological analysis is fundamental to provide reliable and accurate assessments of these changes. Extreme temperature events, such as heatwaves and cold spells, have considerable effects on e.g. health systems, energy consumption and phenological cycles. Their changes in frequency and severity over the last centuries can be studied using daily temperature series from in-situ weather stations. However, these series suffer from external interventions to the measuring stations, such as relocations and modifications to the instruments, and from changes in their surroundings (growing trees, new buildings). These induced changes in the recorded values are not related to climatic events, making the series inhomogeneous and unreliable. With the aim of producing solid temperature databases, several works in the past decades have introduced techniques for the identification of such changes and their correction (*homogenization*). Within this thesis, a new procedure has been developed, taking inspiration from the Quantile Matching method [Trewin, 2013]. This is based on the calculation of different adjustments for average and extreme values and, in this project, has been revisited and modified, introducing new aspects aimed at making it more flexible, more heuristic and more faithful to the originally observed data. The new homogenization is applied to the European Climate Assessment & Dataset (ECA&D), a pan-European dataset providing observations from all European National Meteorological Services. The method is validated with a comparison to acknowledged homogenization methods against a benchmark dataset, proving its robustness and the quality of the results. The homogenized temperature series, thanks to their high reliability, are then analyzed, performing trend analyses focused on the extreme events. Finally the homogenized series are used to create a homogenized version of E-OBS, gridded dataset obtained with the interpolation of ECA&D station data. The homogenized E-OBS is then employed to compare the trends on average and extreme values of the last decades with those simulated in the same period by climate models, which in other studies are used for predictions of future climate under different emission scenarios.

The Introduction (Chapter 1) explains the context in which this thesis is developed. The current state and knowledge about climate change is introduced, followed by the issues implied by the presence of inhomogeneities in temperature series. The aims of the thesis and the expected results and further applications are then exposed in detail.

Chapter 2 describes in detail the statistical bases of the homogenization method. It is composed by a *break detection* that statistically identifies the timing of the occurred inhomogeneities. This is needed since the dataset is too big to handle the available documentation of reported changes occurred to the stations (metadata), which would require a long and labourious process. The second step is the adjustment calculation, developed following previous studies on the technique of *Quantile Matching*. This calculates different adjustments for the daily data according to their position in the temperature distribution, thus handling differently average and extreme values. The reported case studies prove the effectiveness of the routine, showing clear improvement in the quality of the series. The difference in the trends of several indices of minimum and maximum temperature between homogenized and raw series show limited changes in average (between +0.01 and +0.02) and no geographical patterns. Moreover, the trends of homogenized series present a clear improvement of the geographical consistency and a considerable narrowing of their distribution. This proves the increased quality of the dataset.

The work reported in Chapter 3 describes the process of blending of series. This involves, for example, the series of the station in a city centre that was ended and the new one that was started in a close-by rural area or in an airport. The *blending* procedure here described joins these series by concatenating them and by mutually filling their gaps. While on one side this process generates long series, on the other hand the blended series are not necessarily homogeneous. For this reason, the homogenization process exposed in Chapter 2 is adapted and applied to these series. The results of this process is a set of long and homogeneous series that are fundamental for thorough historical climatic inspections. Three case studies help exposing the complexity of the process and its benefits. Finally a trend assessment on the new homogenized blended series has been performed. Similarly to what reported by previous studies, this has revealed steep trends in summer maximum temperatures over the Mediterranean and in winter minimum temperatures in Eastern Europe. The latter is connected with a narrowing of the winter minimum temperatures, while in Central Europe a relevant widening of summer maximum temperatures is observed.

The Quantile Matching homogenization procedure is compared with other methods in Chapter 4. Here two benchmark datasets are generated, concatenating data from homogeneous neighbouring series in the national network of Czech Republic and among series specifically selected within the ECA&D. Two benchmark datasets allow to compare situations with very good data quality and station density (Czech dataset) and with scarcer station density and presence of missing data (European dataset). Three well known methods (DAP, HOM, SPLIDHOM) are evaluated together with the Quantile Matching, making use of a set of metrics, such as Root Mean Square Error, percentage of adjusted data and evaluation of trends in average and extreme values. On the Czech Dataset almost all methods perform very well, proving the quality of their statistical features in favourable conditions. The European Dataset allows to test the robustness of the methods in challenging conditions. Here some methods show difficulties in the homogenization of warm extremes and large percentages of missed adjustments of biased data. The Quantile Matching works very well in both cases, showing good performances, comparable to the results of a prestigious method as SPLIDHOM.

The homogeneous blended series are the bases for a new version of the gridded dataset E-OBS, which is a valuable tool for the validation of climate simulations, such as the ones developed in the frame of the High Resolution Model Intercomparison Project. These models aim at simulating the climate of the period after 1950 and can be compared to observed values to detect how well they reproduced climate variability and trends. Studies of previous versions of climate simulations highlighted underestimations in the trends of (especially warm) extreme events. In Chapter 5 this comparison is performed taking the difference of the trends in average values and in the number of warm (or cold) extreme events above (or below) percentile-based thresholds. The studied models simulate the trends generally well, though they show underestimation of the strong reduction of cold events in Eastern Europe and of the steep increase of warm events in the Mediterranean area.

In the Synthesis (Chapter 6) the obtained results are summarized and discussed, focusing on how they have accomplished the aims of the research. The homogenization method based on the Quantile Matching has shown to work very well on the individual series and on the whole network, reducing the presence of anomalous trends and increasing spatial coherence of the data. The comparison with other methods against a benchmark dataset has validated the quality of the new method and given reliability to the studies performed on the homogenized dataset. These have confirmed the severe warming processes over Europe, highlighting the increased distribution width of summer daily temperatures over Central Europe and the narrowing of the distribution of winter daily temperatures over the Alps and Eastern Europe. Finally, one of the very powerful uses of the results of this thesis has been shown. This is the evaluation of climate simulations against a homogenized gridded dataset, which has allowed to inspect how well the models are able to reproduce the statistical features of the extreme temperature events over the last decades. Moreover, in the Synthesis possible improvements for the homogenization method are exposed together with concluding remarks. The main conclusions of this thesis are the acknowledgement of the high efficiency of the developed method, of the high quality of the obtained dataset and of the important added value that homogenization processes like this provide to climatological analyses and to the solidity of the evidences of climate change.

Riassunto

I cambiamenti che interessano il clima globale ed europeo comportano rilevanti conseguenze su società ed economia. È quindi fondamentale che essi siano compresi e descritti mediante accurate analisi climatologiche. Gli eventi di temperatura estrema, come ondate di calore e di gelo, hanno effetti considerabili, per esempio sul sistema sanitario, sul consumo di energia e sui cicli fenologici. Le variazioni in frequenza ed intensità di questi fenomeni avvenute negli ultimi secoli possono essere studiate utilizzando serie giornaliere di temperatura, registrate da stazioni meteorologiche in loco. Tuttavia, queste serie presentano criticità dovute a interventi sulle stazioni, come trasferimenti e modifiche alla strumentazione, e a variazioni nell'ambiente circostante come crescita di alberi o costruzione di nuovi edifici. Di conseguenza i valori misurati presentano alterazioni non collegate alla variabilità climatica, ma a fattori esterni, rendendo le serie disomogenee. Negli ultimi decenni diversi studi hanno introdotto tecniche per l'identificazione di queste alterazioni e la loro correzione (*omogenizzazione*), con l'intento di produrre affidabili serie di temperatura. Questa tesi presenta un nuovo metodo di omogeneizzazione, basato sul paragone delle distribuzioni di probabilità, ispirato al metodo del Quantile Matching (QM) [Trewin, 2013]. Il QM calcola correzioni differenti per valori medi ed estremi ed è stato qui rivisitato, introducendo nuovi aspetti tali da renderlo più empirico, flessibile e fedele alle osservazioni originali. La nuova tecnica di omogeneizzazione è stata applicata all'European Climate Assessment & Dataset (ECA&D), una raccolta di misurazioni provenienti da tutti i servizi meteorologici nazionali europei. Un raffronto con rinomati metodi di omogeneizzazione, applicati a un dataset di riferimento, ha permesso di verificare la robustezza e la qualità dei risultati ottenuti e di validare il metodo. I trend calcolati sulle serie omogeneizzate sono più attendibili, permettendo una migliore valutazione dei cambiamenti nei fenomeni medi ed estremi. Infine, le serie omogeneizzate sono utilizzate per calcolare una nuova versione degli E-OBS, dataset su griglia ottenuto con l'interpolazione dei dati di stazione di ECA&D. Questo dataset su griglia è impiegato per confrontare i trend di valori medi ed estremi degli ultimi decenni con quelli simulati sullo stesso periodo da modelli climatici, impiegati allo stesso tempo per la previsione del clima dei prossimi decenni.

L'Introduzione (Capitolo 1) descrive il contesto in cui questa tesi si sviluppa. In particolare è riportato lo stato dell'arte delle conoscenze sui cambiamenti climatici, seguito dai problemi implicati dalla presenza di disomogeneità nelle serie di temperatura. Infine sono trattati l'obiettivo di questa tesi, i risultati attesi e le future applicazioni.

Il Capitolo 2 descrive in dettaglio le basi statistiche del metodo di omogeneizzazione. Il primo stadio consiste nel determinare, attraverso strumenti statistici, a che punto della serie temporale si verifica la disomogeneità. Questo è necessario a causa delle dimensioni troppo estese del dataset, che non permettono di gestire le documentazioni relative alle date degli interventi sulle stazioni (metadati), il cui processo richiederebbe un eccessivo investimento di tempo e risorse. Il secondo passaggio include il calcolo delle correzioni, sviluppato secondo la tecnica del Quantile Matching. Tale metodo calcola diverse correzioni per i dati giornalieri a seconda della loro collocazione nella distribuzione delle temperature, permettendo così di gestire differentemente valori medi e valori estremi. I casi-studio riportati provano l'efficacia del metodo. La media continentale dei trend calcolati sui valori medi non presenta rilevanti variazioni rispetto alle serie originali (da +0.01 e +0.02). Tuttavia, la rimozione dei trend anomali comporta una considerevole diminuzione dell'ampiezza della distribuzione dei trend stessi e una maggiore omogeneità della loro distribuzione geografica.

Il Capitolo 3 descrive il processo di integrazione fra stazioni vicine, detto *blending*. Questo accade, per esempio, fra la serie interrotta di una stazione urbana e quella successivamente avviata in una vicina zona rurale o in un aeroporto. Tale procedura consiste nel concatenare due o più serie e nell'usare i loro dati per riempire reciprocamente eventuali periodi scoperti. Le serie così generate hanno il vantaggio di essere lunghe e complete, ma con lo svantaggio di presentare nuove disomogeneità. Per questa ragione, il metodo di omogeneizzazione mostrato nel Capitolo 2 è stato adattato e applicato alla nuova problematica. Il risultato di questo ulteriore processo è quindi un insieme di serie lunghe ed omogenee, essenziali per un accurato studio storico sul clima. La complessità di questo lavoro e i suoi benefici sono presentate mediante tre casi studio. Infine l'analisi dei trend delle nuove serie ha rivelato, in linea con studi precedenti, trend accentuati per le temperature massime estive nell'area Mediterranea e per le minime invernali sull'Europa Orientale. Quest'ultimo risultato è collegato al restringimento della distribuzione delle minime invernali, in contrasto con il rilevante allargamento individuato per le massime estive in Europa Centrale.

Il Quantile Matching è stato poi comparato con altri metodi, come riportato nel Capitolo 4. Due dataset di riferimento hanno permesso di testare i metodi su condizioni ottimali (alta qualità dei dati e densità di stazioni, dataset Ceco) e su condizioni proibitive (inferiore densità di stazioni insieme a più lacune nei dati, dataset Europeo). Il raffronto con tre rinomati metodi (DAP, HOM, SPLIDHOM) è stato effettuato utilizzando degli indicatori come lo Scarto Quadratico Medio (RMSE), la percentuale di dati corretti e il calcolo dei trend di valori medi ed estremi. Tutti i metodi hanno mostrato buoni risultati sul dataset Ceco, rivelando la qualità delle teorie statistiche utilizzate. Al contrario, il dataset Europeo ha permesso di verificare la robustezza dei metodi in condizioni più impegnative. In questo caso, alcuni metodi hanno mostrato criticità nell'omogeneizzazione dei valori estremi collegati agli eventi caldi, insieme ad alte percentuali di mancate correzioni. Il Quantile Matching ha dato risultati soddisfacenti in entrambi i casi, all'altezza di quelli di un metodo ampiamente validato come SPLIDHOM.

Le serie omogenee ottenute dal processo di *blending* sono utilizzate come base per una nuova versione del dataset su gliglia E-OBS, il quale è un prezioso strumento per la validazione di simulazioni climatiche, come quelle sviluppate nel contesto del progetto di comparazione dei modelli ad alta risoluzione (HighResMIP). Tali modelli hanno lo scopo di riprodurre il clima del periodo successivo al 1950 e possono essere confrontati con le osservazioni per esaminare la simulazione della variabilità e dei trendi climatici. Studi sulle precedenti versioni di tali simulazioni hanno evidenziato sottostime nei trend, specialmente degli eventi caldi. La comparazione fra E-OBS e le simulazioni è trattata nel Capitolo 5, in questo caso l'analisi si focalizza sulla differenza dei trend nei valori medi e nel numero di eventi al di sopra (o al di sotto) di soglie calcolate in base ai percentili della distribuzione. I modelli studiati simulano in maniera affidalible i trend di valori medi, tuttavia presentano una sottostima della forte riduzione degli eventi freddi in Europa Orientale e del rapido aumento degli eventi caldi nell'area Mediterranea.

Nelle Conclusioni (Capitolo 6) vengono discussi i risultati ottenuti, valutando come questi rispondano agli obiettivi della ricerca. Il metodo di omogeneizzazione basato sul Quantile Matching ha dimostrato di lavorare in maniera ottimale sulle singole serie e sull'intero dataset, riducendo la presenza di trend anomali e incrementando la coerenza spaziale dei dati. Il raffronto con altri metodi tramite un dataset di riferimento ha permesso di validare la qualità del nuovo metodo e di dare affidabilità agli studi messi in atto sul dataset omogeneizzato. Questi ultimi hanno confermato l'intenso processo di riscaldamento in Europa, evidenziando l'allargamento della distribuzione delle temperature estive in Europa Centrale e il restringimento della distribuzione delle temperature invernali sulle Alpi e in Europa Orientale. Infine, è stato mostrato uno degli usi più significativi dei risultati di questa tesi: la valutazione delle simulazioni climatiche mediante un dataset su gliglia omogeneizzato che ha permesso di testare la capacità dei modelli di riprodurre le statistiche degli eventi di temperatura estrema negli ultimi decenni. Inoltre, sono indicati i possibili miglioramenti da apportare al metodo di omogeneizzazione, assieme a considerazioni finali. Le principali conclusioni tratte in questa tesi sono la validazione dell'alta efficienza del metodo sviluppato, della rilevante qualità del dataset che viene generato e l'importante valore aggiunto che i processi di omogeneizzazione come questo conferiscono alle analisi climatologiche e alla solidità delle evidenze riguardo i cambiamenti climatici.

Contents

Page

Summary vii

Riassunto xi

Contents xv

Chapter 1 Introduction 1

Chapter 2 Homogenization of daily temperature series in the European Climate Assessment & Dataset 19

Chapter 3 Building long homogeneous temperature series across Europe: a new approach for the blending of neighboring series 51

Chapter 4 Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset 75

Chapter 5 Evaluation of trends in extreme temperatures simulated by HighResMIP models across Europe 103

Chapter 6 Synthesis 131

References 143

Acknowledgements 155

About the author 157

Graduate School Certificate 159

Chapter 1

Introduction

1.1 Climate Change

Since the end of 19th century the atmosphere of Planet Earth has experienced a marked warming process attributed to the human activity. The average temperature has raised with a rate of about 0.085°C per decade in the period from 1880 to 2019 and has shown an accelerated rate in the second half of the 20th century [Klein Tank and Können, 2003; Hansen et al., 2010; Simolo et al., 2010; Lawrimore et al., 2011; Jones et al., 2012; Rohde et al., 2013; Delvaux et al., 2018; World Meteorological Organization, 2019]. The trend is confirmed by the fact that each of the four last decades (1980-1989, 1990-1999, 2000-2009 and 2010-2019) has been warmer than all the previous ones [Hartmann et al., 2013; World Meteorological Organization, 2019]. The global average during the last ten years has shown persistence of this tendency: according to WMO [1], NOAA Global Time Series [2], MetOffice/UEA Climate Research Unit [3] and NASA-GISS Surface Temperature Analysis [4], among the top 10 warmest years on records, 7 belong to the last decade with 2019 being the second warmest year ever after 2016, see Figure 1.1.

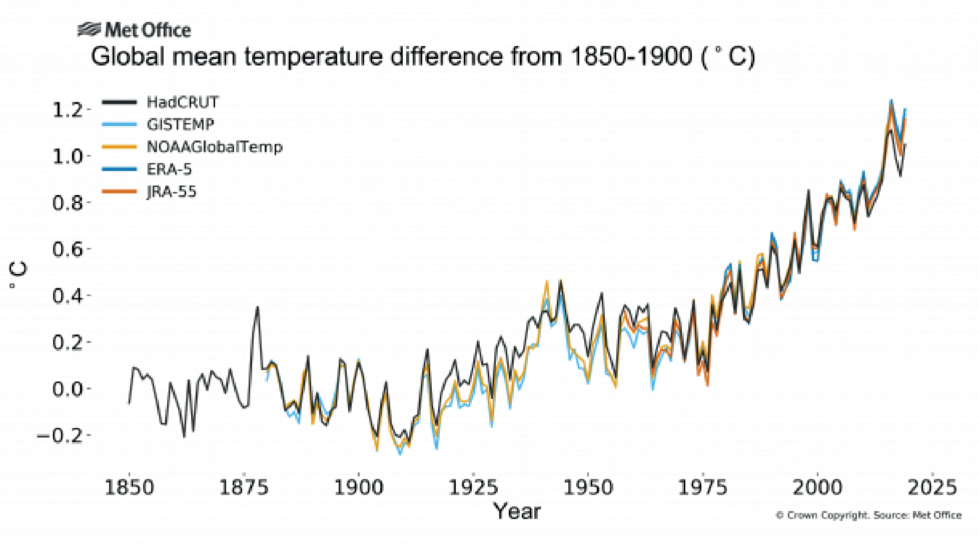


Figure 1.1: *Anomalies of global yearly mean temperature compared to the 1985-1900 mean. Courtesy of WMO, https://public.wmo.int/en/media/press-release/wmo-confirms-2019-second-hottest-year-record*

The natural variability of the climate is not sufficient to describe such changes [Bindoff

[1] https://public.wmo.int/en/media/press-release/wmo-confirms-2019-second-hottest-year-record

[2] https://www.ncdc.noaa.gov/cag/

[3] http://www.cru.uea.ac.uk/

[4] https://data.giss.nasa.gov/gistemp/graphs_v4/

et al., 2013]. There is strong agreement on attributing the reasons of this phenomenon to increased concentrations of greenhouse gases such as carbon dioxide, methane and nitrous dioxide, emitted in the atmosphere by industrial activities. The energy balance of the atmosphere is due to the equilibrium between solar incoming radiation (short-wave) and thermal upward emission of surface and atmosphere: the more solar radiation enters the system, the higher the temperature of the system gets, in order to emit more long-wave radiation and find a balance. The greenhouse gases absorb the upward long-wave radiation emitted by the Earth surface and by the atmosphere itself and re-emit it partially downwards, increasing the incoming energy in the troposphere. The lower layers of the atmosphere respond to this excess raising the average temperature, in order to increase the thermal emission and reach an energy equilibrium.

The warming of temperatures has been observed almost in all places of the surface of the Earth [Hartmann et al., 2013; World Meteorological Organization, 2019]. The magnitude of such phenomenon has been inspected by several studies which have focused on national, regional and global station networks or gridded interpolated datasets. These have highlighted different behaviours and different warming trends, identifying continental areas at high latitudes (including, among them, Europe, Mediterranean and Middle East) as areas affected by a stronger increase of temperatures [Vose et al., 2005; Alexander et al., 2006; Rohde et al., 2013; Donat et al., 2013; World Meteorological Organization, 2019].

While the increase in average temperature has been widely analysed and demonstrated, the dynamics of warm and cold extreme events are currently of great interest due to their strong impact on economy and society. These events are monitored by specific indices that measure, for each year, the number of days that are above or below a certain threshold, which is either a fixed critical value (e.g. below 0°C for frost nights) or a value which relates to local climatic features. The latter option allows analyses that are applicable to any kind of climate and for example make use of the 10th or 90th percentiles, which are those values that are larger than, respectively, 10% and 90% of the considered sample. The percentage of days whose minimum (maximum) temperature is below the 10th percentile is called TN10p (TX10p) or Cold Nights (Day-times), while the percentage of days whose minimum (maximum) temperature is above the 90th percentile is called TN90p (TX90p) or Warm Nights (Day-times) [ETCCDI, 2009].

On a global level, it is observed the predominance of warming trends on all extreme temperature indices. Trends on the extremes of minimum temperatures present larger trends compared to those on maximum temperature, see table 1.1, while for both variables the trends on the warm extremes appear to be steeper (but still within each other's error range).

This indicates that, while the rise of night temperatures is significantly larger than the day-time ones, both experience a increase in extremely warm events which is larger than

Table 1.1: *Global average of trends (1951-2012) of the percentage of cold nights (TN10p), cold day-times (TX10p), warm nights (TN90p) and warm day-times (TX90p). [Donat et al., 2013; Hartmann et al., 2013]*

%/dec	TN	TX
10p	-3.9 ±0.6	-2.5 ±0.7
90p	4.5 ±0.9	2.9 ±1.2

the decrease in the number of extremely cold events.

Such fact doesn't automatically imply that a positive trend in variability is observed. The observed behaviour can also be explained by shift of the distribution of temperatures without changes in the shape [Simolo et al., 2010; Morak et al., 2011; Donat and Alexander, 2012], left panel of Figure 1.2. Nevertheless the frequency of extremes is shown to be more sensitive to the changes in variability than to changes in the mean [Katz and Brown, 1992; Della-Marta et al., 2007], this can also be seen in the diagrams of Figure 1.2.

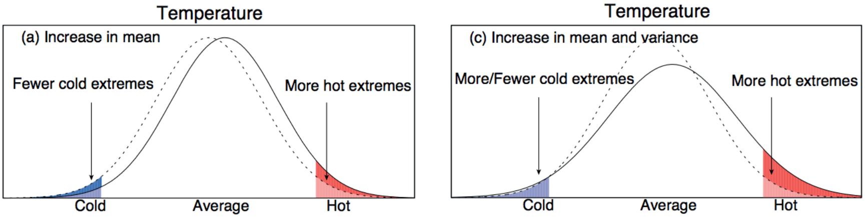


Figure 1.2: Effects on extreme indices due to the increase in the mean and in the variability. *[Cubasch et al., 2013]*

The trends on extreme indices show differences according to the considered areas. Almost all the the regions of the world present trends that can be linked to global warming. Nevertheless wide areas present trends which are significantly larger than on the rest of the globe. Among all, Arctic and Asian areas have shown the largest trends on Cold Nights, while Europe and the Mediterranean have experienced stronger trends on the Warm Days in comparison to the rest of the globe, see Figure 1.3, [Donat et al., 2013]. The latter is directly related to the occurrence of *heatwaves*, which have increased their frequency, length and intensity in the last decades [Scherrer et al., 2005; Della-Marta et al., 2007; Perkins et al., 2012] and whose impact on agriculture, health, infrastructures and economy is considerable [Donadelli et al., 2017].

The present thesis focuses on the European and Mediterranean area, which has been indicated as one of the most sensitive areas to climate change [Giorgi, 2006; Hartmann et al., 2013]. Moreover it has shown considerably large trends in temperatures and in frequency of extreme events such as warm spells and heatwaves [Klein Tank and Können, 2003; Scherrer et al., 2005; Della-Marta et al., 2007; Brown et al., 2008; van Oldenborgh et al., 2009; Efthymiadis et al., 2011; Simolo et al., 2011; Andrade et al., 2012; Perkins et al., 2012; Donat et al., 2013].

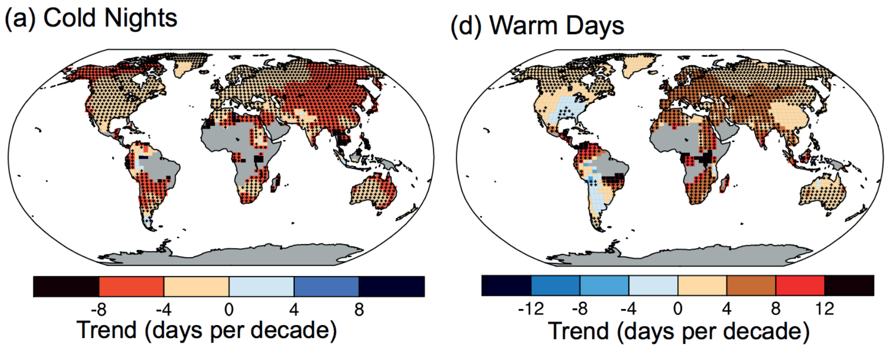


Figure 1.3: Trends in TN10p (left) and TN90p (right) on the global gridded dataset HadEX2. *[Donat et al., 2013]*

Understanding the evolution in frequency and amplitude of extreme temperatures is then fundamental to assess how the climate is changing, to thoroughly communicate to public opinion and policy makers the related risks, to encourage adaptation and increase resilience.

In order to perform this, the best approach is the use of daily temperature measurements from ground-based weather stations. These provide accurate and direct values of the temperature in certain locations and are longer and more solid than series derived from satellite-derived data (which may vary according on the instrument and are the result of inversion processes) and reanalyses data (obtained from data assimilation).

1.2 Temperature measurements and their inhomogeneities

1.2.1 Data collection

The in-situ recording of daily temperatures is a fundamental practice for the production of important data in climate analysis, indispensable for the assessment of magnitude and the speed of the changes in climate. Starting from the second half of 17th century the collection of daily temperature values has slowly spread all over Europe. However only during the 20th century the number of active stations has significantly increased, growing exponentially after World War Two.

The European Climate Assessment and Dataset (ECA&D) [Klein Tank et al., 2002] has operated in the last decades in the collection of series of daily measurements from more than forty-four countries all over European and Mediterranean areas. These series include measurements of temperature and twelve more variables such as precipitation, pressure, wind speed, humidity, cloud cover, solar radiation, etc.. As of beginning of 2020 the ECA&D stores more than 8000 temperature series, having length that ranges from a few years to more than 200 years.

1.2.2 Inhomogeneities

Long series of daily temperatures are fundamental for the inspections on the variations in statistical features (e.g. average, variability, distribution width, etc.) of the temperature distribution [Easterling et al., 1996; Trewin, 2013; Nemec et al., 2013]. However any temporal or spatial analysis is affected by changes occurred to the measuring stations, related to artificial intervention or changes in the natural surroundings [Begert et al., 2005; Thorne et al., 2005; Brunetti et al., 2006; Menne and Williams Jr, 2009]. The relocation of a station from the centre of cities to the airport (or to a closeby rural area) is one of the most common examples of changes that have taken place in the history of a measuring series [Tuomenvirta, 2001; Begert et al., 2005]. This has usually happened to address aviation needs or scientific guidelines. These include for example the regulations of International Civil Aviation Organization (ICAO) that require a meteorological station to be present at the airport and the recommendations of Weather Meteorological Organization (WMO), which advises to place the instrument in a grass field, far from building or trees, in order to ensure reliable measurements. An example of this is the series of Salzburg that initially had been in (various locations of) the centre of the city and in 1939 was relocated to the airport, see Figure 1.4. This is an excellent (and rare) example of a series with exhaustive documentation. Its original version and supplementary information are made available by the Central Institute for Meteorology and Geodynamics (ZAMG), which also produced an homogenized version [Nemec et al., 2013].

The combination of relocations like this with those due to eventual changes of the instrumentation itself (e.g. from manual to automatic thermometers or replacement of the screen [Auchmann and Brönnimann, 2012]) and with changes in the surrounding of the station (such as new building or growing trees, [Ren, 2017; Yosef et al., 2018]) introduce perturbations that significantly alter the natural signal, which is uniquely induced by climate and weather variability [Aguilar et al., 2003; Hartmann et al., 2013]. Such interferences are very common since it is almost inevitable for long-running temperature series to experience any kind of changes [Domonkos, 2011; Kruger and Nxumalo, 2017; Vincent et al., 2018]. As a consequence, in several works more than 70% of the series are found to be inhomogeneous [Brunet et al., 2006; Kuglitsch et al., 2009; Syrakova and Stefanova, 2009; Nemec et al., 2013; Mamara et al., 2014]. These are, most of the times,

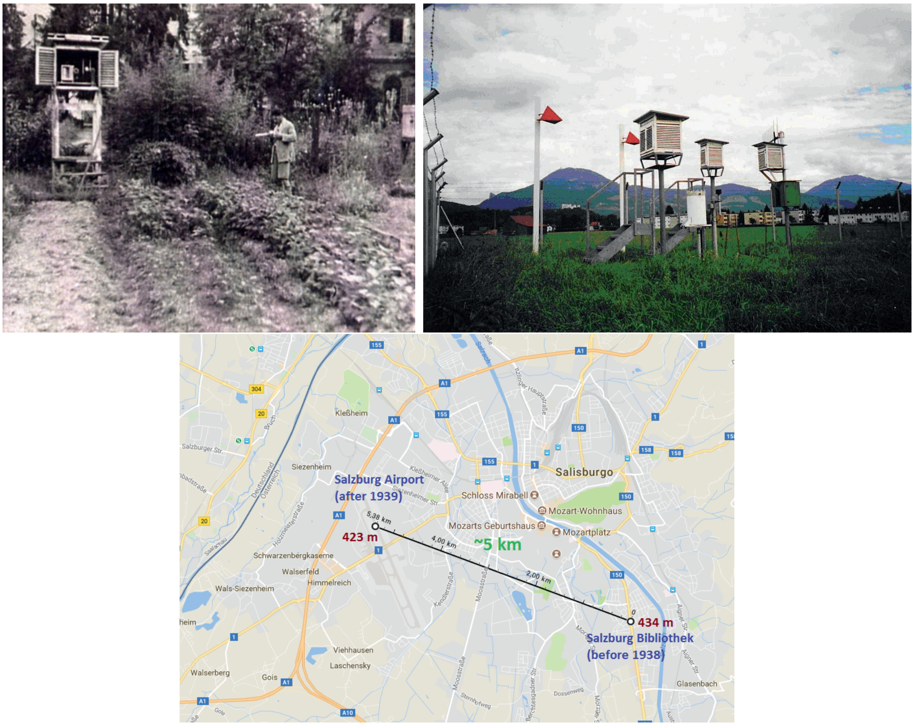


Figure 1.4: *Top left: weather station in the garden of the Library of Salzburg. Top right: weather station in the area of the airport. Courtesy of ZAMG. Bottom: locations of the two stations with distance and elevations. Courtesy of Google Maps.*

related to relocations from urban to rural surroundings [Böhm et al., 2001; Vincent et al., 2012].

In the specific case of Salzburg the relocation to the airport in 1939 induced a sudden cooling bias to the measured values as the measurement equipment were no longer exposed to the warmer conditions which are present in the urban environment, the Urban Heat Island (UHI). This can be distinguished as a step-like change through the warming trend in the annual mean of temperatures of figure 1.5.

1.3 Homogenization: concepts and methods

The presence of inhomogeneities dramatically decreases the accuracy of any analysis based on temperature series. Therefore a process aimed at the adjustment of non-climatic signals in temperature series (i.e. *homogenization*) is essential [Aguilar et al., 2003; Venema et al.,

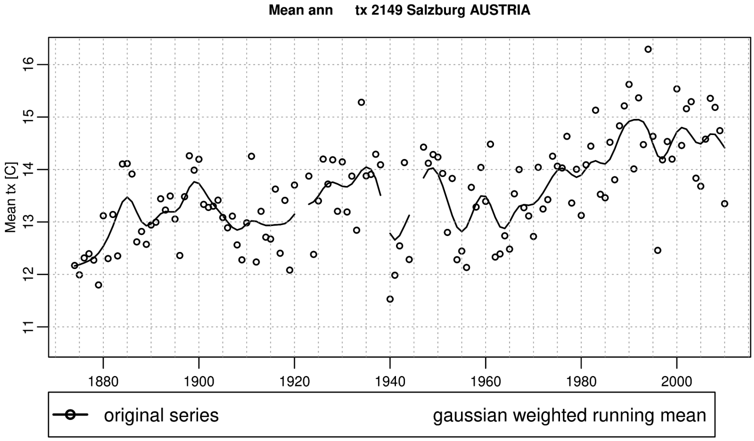


Figure 1.5: *Annual mean maximum temperature at the station of Salzburg. The line represents a weighted running mean, which allows a clearer and less noisy visualization of the tendencies.*

2013]. A climatological study that skips a homogeneity assessment is therefore unreliable and likely to confuse climatic and non-climatic signals, since these often present similar amplitudes [Caussinus and Mestre, 2004].

1.3.1 Statistical Break Detection

The homogenization is composed of two steps: the break detection and the adjustment calculation. The first step can substantially benefits by eventual transcriptions or notes on the changes that may have affected the values of the series, documents known as *metadata*. Such information is extremely precious in the case of manual homogenization, since it allows to establish the temporal location in which the perturbations to the signal (i.e. *breaks*) took place. Nevertheless in most of the cases, especially back in time, this information is not available, not digitized or rather reported in the local language. For this reason many of the methods developed in the last years are completely automated and proceed with statistical routines able to detect the break points [Caussinus and Mestre, 2004]. Furthermore the large size of some networks, such as ECA&D, make it impossible to handle all the metadata, enhancing the need of only automated procedures.

The earliest Break Detection routines (*absolute* methods) were based on analyses of the statistical consistency of the series through time. For example this was done looking for abrupt changes in the mean values, (Standard Normal Homogeneity Test, SNHT, [Alexandersson, 1986]). However, the most recent methods (*relative* methods) consider reference series whose behaviour is a mix of the natural climatic signal and random -

weather - variations. Most of the times the references are chosen among highly correlated neighbouring series, since these have experienced the same climate conditions of the target inhomogeneous series [Aguilar et al., 2003]. This step is particularly important when trends in the temperature are present as these can be misinterpreted by absolute break detection methods as breaks and generate false positives. With the relative method, the assumption is made that the reference series shares the same climatic signal as the target. This makes that, in the case of a homogeneous target, the difference series should have a strictly noisy behaviour, while, when the target is not homogeneous, it is easier to detect a break. Recently the use, as reference, of sets of series has spread (pairwise comparison, [Menne and Williams Jr, 2005; Trewin, 2013]). This allows to identify the climatic signal that is common to multiple neighbouring series [Aguilar et al., 2003; Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006; Venema et al., 2013] and avoids to rely on an individual series [Della-Marta and Wanner, 2006] or on the average of a selected set [Alexandersson and Moberg, 1997; Vincent et al., 2002; Begert et al., 2005; Štěpánek et al., 2009], operation that is sensitive to changes of station availability. Several break detection methods have been developed in the last decades, mostly based on yearly, seasonal or monthly averages [Alexandersson, 1986; Vincent et al., 2002; Caussinus and Mestre, 2004; Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006; Wang et al., 2007; Mestre et al., 2011]. Recent works have compared these methods [Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2013] identifying positive aspects and drawbacks and highlighting the strength of procedures based on agreement among break detections procedure [Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2013], which present lower amount of false positives.

1.3.2 Adjustment calculation methods

Once the break points are detected, each of the portions of the series that is identified by two consecutive breaks needs to be adjusted. This has to be done in order to make them consistent with the latest segment [Aguilar et al., 2003], which is still updated and is likely to be the one with the best quality of measurement recording. The large noise, typical of daily temperature measurements, implies that data can only be adjusted making use of temporal aggregations, usually on a monthly basis. Such approaches allow to inspect the statistical features of the data, such as mean values, variability and parameters related to the distribution.

The adjustment of the series considering uniquely the mean values reduces the impact of the homogenization and is not able to apply adequate corrections to very cold and very warm values. This is due to the fact that relocations, changes in instruments and the other inhomogeneity sources don't induce uniform biases to the data [Tuomenvirta, 2001; Della-Marta and Wanner, 2006]. Each inhomogeneity generator produces different effects on the values, these depend on external factors, such as radiation, cloud cover, wind strength and direction [Brandsma and Van der Meulen, 2008; Brugnara et al., 2016] and are not constant throughout the year [Böhm et al., 2010]. For example in a hot sunny July day a station in the city measures the temperature in an environment that is warmed up by solar radiation and by the heat coming from buildings, asphalt, etc.. The last factors are (partially) absent in the airport or the country side, thus a relatively large difference between the two sites is expected. In a less warm, cloudy and windy day of the same month, where the lower atmosphere is horizontally mixed and the incoming solar radiation is less dominant, the effect related to the urban environment (Urban Heat Island Effect) is expected to be lower. For these reasons adjusting the two cases with same factors would be not accurate.

Therefore the monthly values, obtained e.g. as difference between the averages before and after the break, need to be *projected* on a daily resolution, some works have implemented this by reconstructing the seasonal cycle [Vincent et al., 2002; Brunetti et al., 2006] (i.e. the adjustment depends on the calendar day). Nevertheless, these approaches don't account for the variability or the shape of the temperature distribution [Mestre et al., 2011]. For this reason recently some methods have started taking into consideration the position of the data in the distribution of the inspected month, distinguishing among cold, average and warm events. In order to do this, the two samples of measurements (before and after the break) can be split into bins according to their measured value. These bins usually are equally large and include an established fraction or percentage of the whole dataset, for this reason they are called quantiles or percentiles. The sequences of quantiles of the part before the break (the one that needs to be adjusted) can be compared with other sequences obtained from surrounding homogeneous stations, which are used to calculate the expected values for that series. Such comparison can be performed with sophisticated techniques (non-linear regressions, cubic smoothing splines, etc.) [Della-Marta and Wanner, 2006; Mestre et al., 2011], allowing to account for both the seasonal cycle and the higher moments.

Recently a more empirical approach has been elaborated: the *quantile matching*. The analysis is still based on quantile binning, though it avoids any kind of parametrization or regression, relying on simple operations that keep values closer to the original signal [Štěpánek et al., 2013; Trewin, 2013]. Figure 1.6 shows, for the case of Salzburg, the pdf of the data of the maximum temperatures in the month of January for the 20 years before and the 20 years after the relocation to the airport. Here it's possible to see how the distance between the two distributions is larger for cold values than for warm values. This is more evident in the lower panel where the quantile sequences (5th ,10th ,15th , etc.) of the two distributions are displayed.

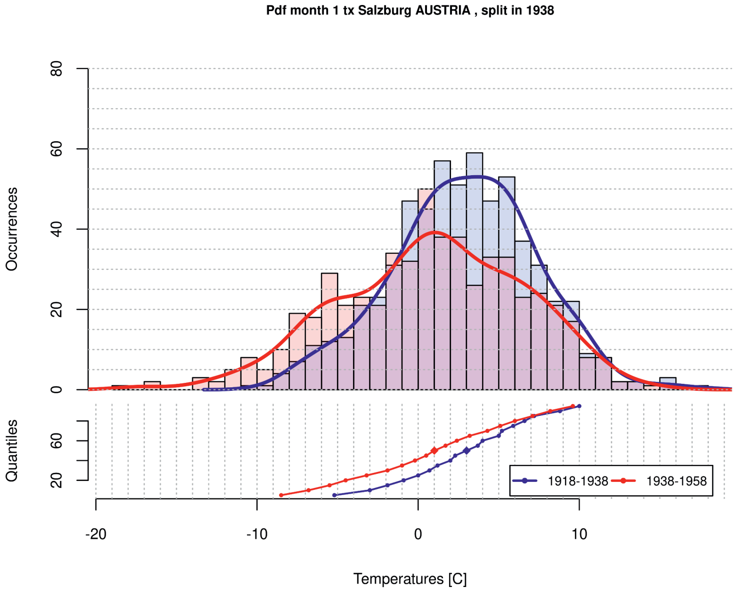


Figure 1.6: *Top: Probability density function of the measurements of maximum temperatures in January in Salzburg. Blue (Red) indicates the portion of 20 years before (after) the break in 1938. Bottom: Sequences of quantiles (5^{th} , 10^{th} , 15^{th} ...) for the same intervals of top panel.*

1.3.3 Application to composite series

In some cases the available metadata (although often insufficient) permits to identify those series that are related to different locations and to split them in more series. After a preliminary check for further inhomogeneities, which is always needed, these shorter series can be employed for the creation of long composite series. Longer series allow more reliable statistical examinations and, by consequence, more solid historical analysis of climate change in the considered locations. The composition can be performed with simple concatenation of series, i.e. series in the airport starts when the series in the city stops, or with a mutual gap filling, i.e. any missing data in the airport series is filled with data from the city, if available. Nevertheless, these processes introduce further inhomogeneities, compromising the reliability of the long series. While the concatenation generates a unique break date, the mutual gap filling creates a *blending* of data from different sources that requires particular care [Trewin and Trevitt, 1996; Tuomenvirta, 2001; Brunet et al., 2006; Böhm et al., 2010; Rahimzadeh and Nassaji Zavareh, 2014; Vincent et al., 2018]. Therefore the homogenization of the composite series is difficult for the particular cases that can be present. However, this step is fundamental to turn the long composite series into long and homogeneous ones, powerful tools for accurate climatological analyses [Peterson et al., 1998; Aguilar et al., 2003]

1.4 Aims of the Thesis

The present thesis has the purpose to provide a significant contribution to the improvement of reliability of climatological analyses of extreme events based on in-situ daily temperature measurements. These series of daily data are of primary importance, compared to reanalyses and satellite data. This is because of their longer duration and to the fact of being a direct measure of the local temperature, not needing inversion or data assimilation procedures. Hence, they are essential for a thorough description of the change and the variability of the climate in the last centuries. The assessment of their quality and homogeneity is mandatory, in order to guarantee reliability and robustness. The removal of non-climatic signals is a fundamental part of this process.

The starting goal of this thesis is to plan and develop a homogenization routine for the ECA&D temperature dataset. Such method can take inspiration from some of the best aspects of existing methods and improve them with innovative solutions. The large size of the European Climate Assessment & Dataset and the impossibility to process the metadata (stored in the shape of comments) require the development of a *completely automated* procedure. An important corner stone for automated procedures is to avoid the overcorrection of the data. This implies the need to keep the values and the statistical features of the series as close as possible to the original version. Such *conservative* approach includes the requirement of a break detection method with a low rate of false positives. The procedures based on the agreement of a set of methods [Kuglitsch et al., 2012; Venema et al., 2013] present such quality. At the same time the adjustment calculation method is required to manage the data with care, preserving the statistical and physical features of average and extreme values. This last requirement is a fundamental aspect of the work of this thesis, due to the important role of extreme temperatures events on e.g. agriculture, health. [Donadelli et al., 2017]

For this reason the procedure developed in this thesis takes inspiration from previous works as the study of Trewin [2013], which introduced a new *heuristic* approach to the homogenization methods based on the analysis of the distribution of temperatures. The absence of linear or non-linear regressions (used by Della-Marta and Wanner [2006]; Mestre et al. [2011]), averages of values of surrounding stations [Štěpánek et al., 2013] or fitting to predefined functions [Menne and Williams Jr, 2009] permits to focus on the comparison of the distributions of the original data. This is a key point for this work, since it falls within the scope of the implementation of a conservative approach. Such pragmatic avenue is expected to convey high *flexibility* to the statistical aspects of the method.

This is fundamental due to the wide spectrum of inhomogeneity sources, whose signals take many forms. Furthermore the need of flexibility involves aspects as the selection of reference series. This step of the procedure has to be able to work in the great variety of station density that the ECA&D presents. For example, in some areas and periods, one series might have a very low amount of neighbouring (potentially usable as reference) series. Such conditions require careful constraints on the references selection. Indeed, the choice has on one side to preserve the statistical robustness, selecting an adequate number of references and, on the other side, avoiding the involvement of low correlated series, which would carry misleading signals [Domonkos, 2011].

Once the original series have been homogenized, it is important to check the presence of series which lay in neighbouring areas and to combine them into longer series. The use of long lasting records brings considerable benefits to climatological analyses, since they allow valuable historical perspectives. Thus, a further goal of this thesis is the application of a process of composition of series, inspired to the blending process already developed within the ECA&D team. Nevertheless, since this process is based on mutual donation of data for the filling of gaps, one value related to an original series can be present in more than one blended series. For this reason a procedure of duplicate removal is needed, in order to have that value appearing in only one blended series. Finally, the inhomogeneities generated by the gathering data related to different locations need to be adjusted, applying an adapted version of the homogenization procedure developed within this project.

For the requirements listed above, this thesis contributes significantly in the scientific debate about the development of automated homogenization processes for daily temperature series. The presence of important aspects such as being conservative, heuristic and flexible are expected to guarantee a powerful and reliable result. Thus the resulting homogenized series are precious for historical and climatological studies on average and extreme values.

In order to assess these benefits, this thesis aims at verifying how much and in which way the homogenization modifies the statistical behaviours of a temperature dataset. This can be done performing comparison with the original versions of the data. A crucial task of this step is to prove that this homogenization process does not have the purpose of enhancing the observed warming trends or hiding tendency with opposite sign. Hereafter, a thorough validation of a homogenization methods needs to include the comparison with other homogenization methods against a solid benchmark, making use of clear and transparent metrics.

These steps provide solidity to the historical analyses that are performed on the homogenized dataset. These studies, which are part of the aims of this thesis, intend to describe the changes in average temperatures and extreme events over Europe in the last centuries.

At this point, the assessment of the local characteristics of climate change is of primary importance, including inferences the possible reasons of the observed trends.

The benefits of working with a homogenized temperature dataset are several and include the improvement of E-OBS, gridded dataset based on the interpolation station data of ECA&D. The wide use of E-OBS in several fields (biology, energy market, health, etc.) makes it a powerful product, whose use is also possible for the validation of climate models. Providing a powerful example of the benefits of the use of homogenized data for other disciplines is the final aim of this thesis.

1.5 Expected results and benefits

1.5.1 Effects of the homogenization

The application of homogenization procedure and the analysis of its effects on national networks or individual stations has been the subject of several studies in the last decades. The large amount of inhomogeneities due to relocations often introduce cooling signals to the new measurements, similarly to what observed for the case of Salzburg.

In these cases, as expected, the predominance of negative correcting factors (to be applied to the earlier portion of the series) has been reported in several analyses [Tuomenvirta, 2001; Böhm et al., 2001; Syrakova and Stefanova, 2009; Lawrimore et al., 2011; Vincent et al., 2012; Nemec et al., 2013], see Figure 1.7. The lowering of the earlier values compensates the cooling effect of the inhomogeneities, implying, by consequence, a warmer trend in the homogenized series. Nevertheless, these examples haven't represented the totality of the adjusted series [Brunet et al., 2006; Nemec et al., 2013; Mamara et al., 2014; Yosef et al., 2018; Squintu et al., 2019] and in some cases positive adjustments have been found to be dominant [Kuglitsch et al., 2009; Osadchyi et al., 2018], thus generating less warm trends.

The reduction of standard deviation (due to the removal of step-like changes) and the increase of mutual correlation that is observed Mamara et al. [2014] have confirmed the trustworthy of the homogenized dataset. The improved features of the networks have also been seen in the better spatial consistency of the trends [Syrakova and Stefanova, 2009; Hannart et al., 2014; Delvaux et al., 2018; Osadchyi et al., 2018; Yosef et al., 2018; Fioravanti et al., 2019] and in the removal of the anomalous ones [Vincent et al., 2002; Mamara et al., 2014; Pérez-Zanón et al., 2015; Squintu et al., 2019] as can be seen in Figure 1.8.

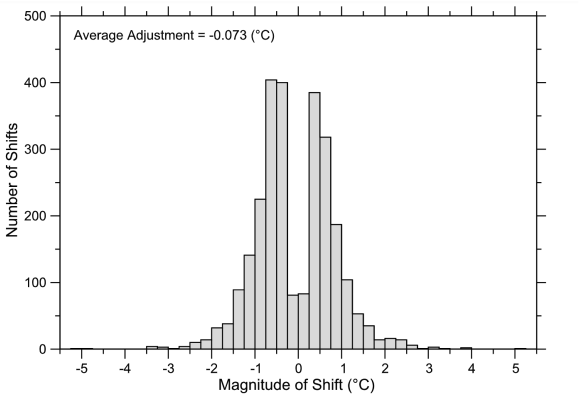


Figure 1.7: *Histogram of adjustments of daily series in the global dataset (escluding US) used by [Lawrimore et al., 2011], calculated with the method by [Menne and Williams Jr, 2009]*

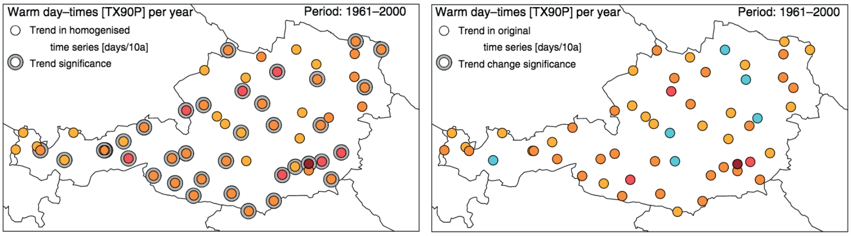

Warm day–times [TX90P] per year
Period: 1961–2000
○ Trend in homogenised time series [days/10a]
◎ Trend significance
Warm day–times [TX90P] per year
Period: 1961–2000
○ Trend in original time series [days/10a]
◎ Trend change significance

Figure 1.8: *Trends on TX90p over Austria after (left) and before (right) the homogenization. [Nemec et al., 2013]*

1.5.2 Comparison of homogenization methods

Different methods applied to the same raw data are expected to produce homogenized dataset with slightly or significant different features [Pérez-Zanón et al., 2015; Li et al., 2016; Fioravanti et al., 2019]. Consequently, it is reasonable to wonder which homogenization method to choose when approaching an inhomogeneous dataset. Important studies have compared homogenization procedures (manual, semi-automated and automatic), mainly focusing on the break detection accuracy and on the adjustment calculation on a monthly level [Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2016]. On the other side several works have performed comparisons of homogenization methods with the intent of validating a newly developed one [Caussinus and Mestre, 2004; Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006; Mestre et al., 2011; Trewin, 2013].

A thorough comparison must be transparent and objective, this is possible by using benchmark datasets and clear metrics. Benchmark series are homogeneous series which are intentionally perturbed, generating inhomogeneous series (*test*) whose truth is known. This can be done introducing missing values, outliers, trends, noise and inhomogeneities [Menne and Williams Jr, 2005; Mestre et al., 2011; Williams et al., 2012; Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2016] with known frequency and amplitude, determined within ad-hoc studies. However, such approaches don't reproduce all the possible signals that can be generated by real inhomogeneities [Vincent et al., 2018]. For this reason in the last years the generation of test series using observed data has become more common [Trewin, 2013]. This has usually been performed by concatenating portion of series related to neighbouring series, so that all the statistical features of a real relocation (plus eventually other inhomogeneity sources) can be simulated.

Once the benchmark dataset has been generated, the homogenization methods have been applied to the tests, producing a set of homogenized versions. These have been compared to the benchmarks with clear metrics such as the root mean square error [Mestre et al., 2011; Venema et al., 2013; Domonkos, 2013; Trewin, 2013; Domonkos and Coll, 2017; Gubler et al., 2017; Vincent et al., 2018]. Finally the differences between the trends in the benchmarks and in the homogenized versions (in average and extreme values) has allowed to evaluate if and how the main climatic characteristics of the benchmarks are reproduced. The comparison of homogenization methods are powerful tools that stimulate the scientific debate in the field and push researchers towards the improvement of the methods themselves.

1.5.3 Use of homogenized datasets in climatological studies

Homogeneous temperature series are a precious tool for the creation of gridded datasets. In the particular case of this dissertation, the homogenized blended series of ECA&D are used for the calculation of a new version of E-OBS [Haylock et al., 2008; Cornes et al., 2018], a high resolution (0.1°) gridded dataset that covers almost all Europe and Mediterranean from 1950 to present days. The lower uncertainty of the values and the higher spatial consistency make it more reliable for the studies of any field, from climatology itself, to health, agriculture and infrastructure-related studies. A particularly important application is the use of homogenized gridded datasets for the validation of climate simulations [Bhend and Whetton, 2013; Flato et al., 2014]. These are models that run on the past decades taking historical observed values (land use, greenhouse gases aerosol concentration, volcanic eruptions, solar radiation, etc.) as boundary conditions and forcings. The evaluation of the biases and the comparison of trends on average and extreme values allow to assess the accuracy of the models. Evaluations performed in the last years, especially after the release of the Climate Model Intercomparison Project 5 (CMIP5) [Taylor et al., 2012], have shown that the models reproduce well the trends of average values [Sillmann et al., 2013; Flato et al., 2014]. Nevertheless a relevant underestimation of the trends of extreme warm values has been observed over Western Europe and Mediterranean [Min et al., 2013; Flato et al., 2014]. Since Europe is one of the most sensitive areas to climate change [Giorgi, 2006; Brown et al., 2008; Hartmann et al., 2013], the new homogenized E-OBS have a high potential in determining how good the models have simulated the climate and, consequently, in speculating on how good they may predict the changes in climate in the next decades.

Among all the further possible uses, homogenized series are also a precious tool for the evaluation of small scale phenomena, like those related to boundary layer dynamics. In particular, models like Weather Research and Forecasting System (WRF) can benefit from the presence of series that carry exclusively climatic signal and can compare the reproduced Urban Heat Island with the adjustments calculated within the homogenization processes.

More uses of homogenized temperature can be implemented. This happens for health related studies (e.g. impacts of heat waves or cold spells), phenological analyses, biological inspections (for example analyzing the relation between temperature and vectors of diseases, as ticks) and energy demanding analyses (for heating or cooling). Finally even within Earth and Climate sciences there are high potentialities that come from the use of homogenized data, such as the inspection of the influences of temperature extremes on the abundance of particular compounds (e.g. NO_2) and, as mentioned above, modelling improvement or boundary layer analyses.

1.6 Structure of the thesis

The thesis is composed by this Introduction, the following four Chapters and a final Synthesis.

The detailed mathematical and statistical bases of the Quantile Matching methods are displayed in Chapter 2. Here all the steps are explained and the effects on individual relevant case studies are shown. Furthermore an outlook on the trends of temperature series before and after the homogenization allows to evaluate the consequences on the trend distribution and the increased spatial consistency of the dataset.

Chapter 3 presents the composition of blended series with a sophisticated step for the removal of the duplicate data generated by the mutual gap filling. After this, the application of Quantile Matching to the blended series is presented, together with three peculiar case studies. Important section of this chapter is a trend assessment over the whole continent, with particular attention on the trends of extreme events and on the difference between them.

The Quantile Matching is compared with other automated homogenization methods (DAP [Štěpánek et al., 2013], HOM [Della-Marta and Wanner, 2006] and SPLIDHOM [Mestre et al., 2011]) in Chapter 4. Here a new procedure for the generation of a benchmark dataset, based only on observed data, is described. The applications of the homogenization methods to this benchmark dataset have been evaluated with transparent metrics and with new indicators that allow to evaluate how accurately the trends are reconstructed.

Chapter 5 shows how the E-OBS gridded dataset, obtained by interpolating [Haylock et al., 2008; Cornes et al., 2018] the homogenized blended series, can be used to validate the climate projections of the model developed in the frame of High Resolution Model Intercomparison Project. This is done by looking at trends in average and extreme values and introducing an index that is able to compare the new higher resolution models with the performances on trends of the lower resolution version.

Finally in the Synthesis, Chapter 6, the highlights and the main conclusions are outlined. In particular here the inspection will focus on how the obtained results fit the aims of the thesis. Furthermore this chapter will introduce the new issues that have arisen during the development of the project and the possible improvements that can be implemented to further increase the value of the homogenization procedure and of its results.

Chapter 2

Homogenization of daily temperature series in the European Climate Assessment & Dataset

Abstract

The daily maximum and minimum temperature series of the European Climate Assessment & Dataset are homogenized using the quantile matching approach. As the dataset is large and the detail of metadata is generally missing, an automated method locates breaks in the series based on a comparison with surrounding series and applies adjustments which are estimated using homogeneous segments of surrounding series as reference. A total of 6,500 series have been processed and after removing duplicates and short series, about 2,100 series have been adjusted. Finally, the effect of the homogenization of daily maximum and minimum temperature on trend estimation is shown to produce a much more spatially homogeneous and then plausible picture.

This chapter was originally published as:
Squintu, A. A., van der Schrier, G., Brugnara, Y., and Klein Tank, A. (2019). Homogenization of daily temperature series in the european climate assessment & dataset. *International Journal of Climatology*, 39(3):1243–1261
https://doi.org/10.1002/joc.5874

2.1 Introduction

Modifications to meteorological stations, such as relocation, replacement of the instrument, recalibration, new buildings in the neighborhood or growth of vegetation in the proximity, alter temperature measurements and introduce biases in the observational records that do not relate to weather and climate [Aguilar et al., 2003; Hartmann et al., 2013]. The analysis of climatic variability and climatic change requires homogeneous temperature series [Peterson et al., 1998]: these series do not confuse the climatic signal with artificial biases which are present in non-homogenous series [Begert et al., 2005; Thorne et al., 2005; Brunetti et al., 2006; Menne and Williams Jr, 2009]. Prior to climate analyses, actions are required aimed at the removal of step-like or gradual changes related to these non-climatic effects in observational records [Caussinus and Mestre, 2004].

The registration of activities on meteorological stations in the metadata keeps track of these changes and sufficiently detailed metadata allow a precise temporal localization of the breaks. Unfortunately the availability of metadata is often low, especially further back in time, and doesn't cover the whole set of inhomogenities that affect the measurements [Caussinus and Mestre, 2004]. This implies that break-detection based on metadata only is not possible for many datasets, even though this approach is regarded as most accurate and reliable. This argument, and the sheer size of a dataset, motivates the use of an automated homogenization procedure [Caussinus and Mestre, 2004].

The aim of this study is to develop a pan-European homogeneous dataset of daily maximum and minimum temperature using such an automated homogenization procedure. It will use a recent agreement-based system to detect breaks [Kuglitsch et al., 2012] and the quantile matching technique [Trewin, 2013] in combination with a pairwise-comparison [Menne and Williams Jr, 2005] approach to determine adjustments. The elements in this approach are introduced below.

Automated homogenization procedures consists of two steps: break detection and adjustment calculation (which follow - or are integrated with - a quality check procedure) [Alexandersson, 1986; Caussinus and Mestre, 2004]. These have been focusing mainly on the detection of breaks in the monthly, seasonal or annual values and use statistical tests accompanied with penalizing functions [Alexandersson, 1986; Caussinus and Mestre, 2004; Menne and Williams Jr, 2005; Wang et al., 2007] or inspections on autocorrelation of residuals [Vincent, 1998]. Recent comparisons [Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2013] have pointed out advantages and drawbacks of the most common systems. Procedures that look for an agreement among methods (e.g. [Kuglitsch et al., 2012]) go one step further and take benefits from the reduced uncertainty in break location by looking for consensus.

Homogenization of annual or monthly averages does not automatically imply a homogenization of higher-order moments [Trewin, 2013] since the processes that generates inhomogeneities on daily data-sets are non-linear, i.e. introduced inhomogeneities to the temperature measurements depend, not in a linear way, on the temperature itself [Della-Marta and Wanner, 2006] and external factors as cloud cover, wind strength and direction can modify extreme daily values differently than the averaged conditions [Brandsma and Van der Meulen, 2008].

Some methods homogenize daily records simply by interpolating monthly adjustment to a daily resolution via a polynomial [Vincent et al., 2002] or trigonometric regressions [Brunetti et al., 2006]. While this approach assures that daily adjusted values reflect the same temporal behaviour as those observed in the monthly series [Vincent et al., 2002], adjustments for the higher-order moments are not guaranteed [Mestre et al., 2011]. A more advanced set of methods considers the temperature distribution which, split into quantile bins, is compared with expected values obtained from surrounding stations. A non-linear regression or cubic smoothing splines [Mestre et al., 2011] are used for the calculation of the correcting factors. Finally a method not based on model parameterization or regressions is the quantile matching, which compares quantiles of the distributions of measurements before and after the break and calculates adjustments by requiring similarity between these distributions.

Ideally, adjustments are made by comparing measurements from the original and the disturbed situation for overlapping periods [World Meteorological Organization, 2011]. The difference between these records eliminates the background climatic signal and highlights the effects of e.g. the change in location. When such parallel measurements are not available, the most reliable source of information about the climate background is the net of neighbouring stations being exposed to the same climatic conditions as the target series [Aguilar et al., 2003; Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006; Venema et al., 2013].

Approaches to construct the reference series are weighted (or simple) averages of surrounding recorded anomalies [Alexandersson and Moberg, 1997; Vincent et al., 2002; Begert et al., 2005; Štěpánek et al., 2009], using a high-correlated homogeneous series [Della-Marta and Wanner, 2006] or performing the *pairwise comparison* [Menne and Williams Jr, 2005; Trewin, 2013]. The calculation of an averaged series may incorporate the inhomogeneities of neighbouring series into the reference series [Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006], might get misleading features from uncorrelated series and can be affected by the change in number of contributing series, introducing strong changes of mean and variance in the reference [Brunetti et al., 2006], thus compromising the ability to represent statistical features of the climate background [Caussinus and Mestre, 2004]. On the other hand, while the use of single series allows the analysis to be independent from the changes in data availability, this approach is risky since it relies totally on a series whose quality might not be certain and whose climatic features might not be consistent with the target series. Isolation of the artificial signal with pairwise comparison, where each reference series provides an estimate of the adjustment of the target series, is shown to be more robust at detection undocumented changes [Menne and Williams Jr, 2009] and more reliable for the calculation of estimates of the adjusting factors [Trewin, 2013].

The study is organised as follows: Section 2 introduces the data set and the methods, Section 3 shows the results, a few case studies and the effects of the homogenization on trends in temperature. The study is discussed and concluded in Section 4.

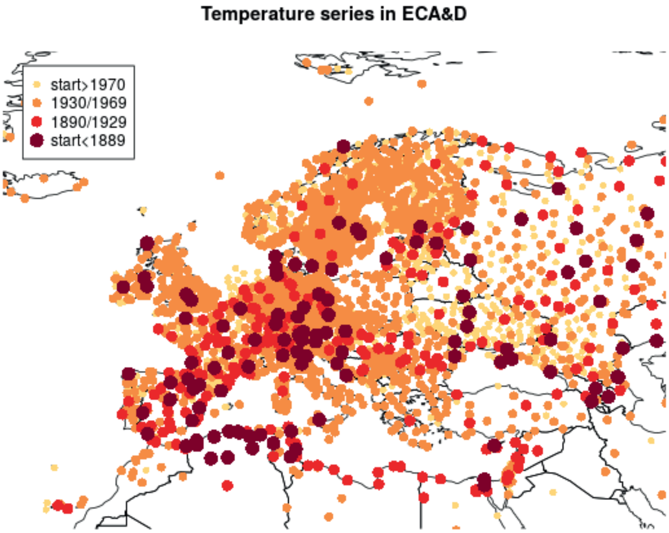
2.2 Data & Methods

2.2.1 ECA&D Dataset

The European Climate Assessment & Dataset (ECA& D, [Klein Tank et al., 2002] [Klok and Klein Tank, 2008] is a collection of daily station observations of currently 12 elements and contains at the time of writing (July 2018) data from nearly 11100 European stations (more than 7500 with temperature measurements) and is gradually expanding. ECA&D contains more than 200 temperature series starting before 1900, but a strong increase in the number of series is found in the 1950s. Data from the station network at ECA&D is updated on a monthly basis using data kindly provided directly by the National Meteorological and Hydrological Services (NMHSs), individual researchers affiliated with a university, global data centres like the National Centers for Environmental Information (Asheville, US) or the synoptic messages from the NMHSs delivered through the Global Telecommunication System [World Meteorological Organization, 2007].

Data coverage varies depending on the countries and on the time coverage (Figure 3.1. Early series (start before 1890) are well distributed in western Europe with the exception of southern Italy and northern Scandinavia. Further data-rescue work is in progress for the improvement of data coverage in low density areas and periods.

The quality check procedure in ECA&D is documented elsewhere [ECA&D Project Team, 2012] and at the time of writing simply insists on consistency between maximum and minimum temperature, does not allow more than 5 repeating values and flags data when the difference from the climatological value exceeds five times the standard deviation. A more sophisticated quality check procedures for ECA&D is currently developed, as an evolution of Štěpánek et al. [2009], based on spatial consistency of measurements in cooperation with Global Change Research Institute (Brno,CZ).



Temperature series in ECA&D

Figure 2.1: *Minimum and maximum temperature series in ECA&D. Colour code indicates the start of the series.*

2.2.2 Break Detection

The detection of breaks is done using a completely automated procedure which is blind to metadata. The break detection method is inspired by the approach of Kuglitsch et al. [2012], which seeks agreement (i.e. common detected timing of breaks) of two out of three common break detection methods (Prodige: [Caussinus and Mestre, 2004], RHtest: [Wang et al., 2007] and GAHMDI: [Toreti et al., 2012]. The main difference with the Kuglitsch et al. (2012) approach is that both the selection of reference series and the combination of the three detection methods have been automated. Both are performed separately on annual and winter/summer half means of standardized differences between candidate and a maximum of eight reference series are used, basing the selection on completeness, correlation of annual averages (minimum of 0.6), and distance (maximum of 1000 km). At least three reference series must confirm a breakpoint in any of the temporal aggregations in a pairwise approach.

The breakpoints are detected at annual resolution. Breakpoints detected in adjacent years by different methods, reference series, or temporal aggregations, are considered the same breakpoint.

Simultaneous changes made to the measurement networks at a national scale are difficult to detect simply because surrounding reference series will suffer from the same resulting break. For these breaks documented metadata is required.

2.2.3 Calculation of Adjustments: Quantile Matching

The adjustment of daily temperature is inspired by the work of Trewin [2013] and is based on a quantile matching algorithm which compares the probability density distribution of temperature before and after the considered break, not taking into account metadata. By making use of a set of homogeneous references series, the climatic signal is accounted for and the assumption is made that the difference series between the candidate and the reference (in their homogeneous sub-periods) is random noise.

The adjustment process targets each series individually. The break detection produces a sequence $(t_1, t_2, t_3, ..., t_n)$ of the timing of the detected breaks in the candidate series (from the most recent to the earliest). Following these breaks, the candidate is divided into $n + 1$ sub-series

$$\mathbf{S}_0(\mathrm{t}|t_1 < \mathrm{t}),\ \mathbf{S}_1(\mathrm{t}|t_2 < \mathrm{t} < t_1), \text{ etc.,} \tag{2.1}$$

which are homogeneous by definition [Caussinus and Mestre, 2004]. These segments will be considered independently during the following steps of the process. Segments shorter than 5 years are not adjusted because of insufficient length required for a robust calculation of quantiles.

Reference selection and use

The references are selected from a box of 6° centered on the candidate station and with an elevation difference smaller than 500 meters. For high-elevation stations ($\geq$ 1000m), this threshold is changed to find neighbouring stations within half the elevation of the candidate station (which increases the number of reference series for mountain stations). Among the series that fulfill these requirements, in case of densely covered areas, the set union of the 40 longest ones and the 20 starting earliest are chosen.

With the same splitting procedure used for the candidate series, the results of break detection are used to divide the reference series into homogeneous sub-series. Only the sub-series with at least 5 years of overlap with both segments of the candidate (i.e. at either side of the break) are selected. This constraint helps to avoid series that have breaks in the same period to be references to each other (e.g. in case of simultaneous breaks in a national network). Since the presence of a trend on temperatures is likely to happen, the maximum length of the sub-series used for calculating the adjustments is set to 20 years, so that the changes in time of the moments don't alter the shape of the distribution, e.g. making it broader.

For each reference sub-series, the daily raw correlation with the segment of the candidate after the break is calculated. In order to limit the computational time but simultaneously preserve statistical significance, out of the set of reference series, the 18 series with the highest correlation are chosen, provided they have correlations higher than 0.75. Note that this threshold is higher than earlier suggested [Domonkos, 2013]. Figure 2.2 illustrates the selection of homogeneous references for a detected break in the candidate series.

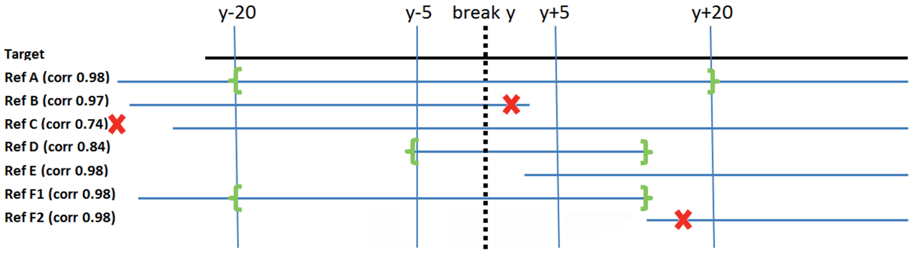


Figure 2.2: *Schematic diagram of reference selection where the sections between green curly brackets are used and sections marked with a red cross are not. Reference series with correlation below 0.75 are discarded, as well as series with overlap shorter than 5 years or data recorded more than 20 years before or after the break.*

In areas with a sparse network, which are common for the early periods, the number of available references may be low. In these cases non-split (and thus inhomogeneous) series (up to 5 in total, meeting the correlation, geographical and temporal overlapping requirements) are added to the reference set, avoiding stations having a sub-segment already selected. In any case a minimum of 3 references series is required for the procedure to be performed, otherwise the candidate is temporarily discarded. The larger size and density of the station network of ECA&D in comparison to the Australian dataset of Trewin [2013] makes that the availability of reference series is higher for the European situation.

Note that the selection of reference series for the adjustment calculation is different than the one used for the break detection.

Calculation of quantile-based adjustments

Adjustment calculation has been developed taking inspiration from [Trewin, 2013]. It is performed backwards considering the breaks from t_1 to t_n successively. The segment of the candidate series after the considered break (t_i) is termed the *basis* series ($\mathbf{B}$), while the *segment* ($\mathbf{S}_i$, where i=1,2,...,n) immediately before it is to be adjusted. For each reference $\mathbf{R}_j$, we define $\mathbf{R}_j^{\text{aft}}$ and $\mathbf{R}_j^{\text{bef}}$ as the portion of the reference after and before the break t_i.

The quantile-based adjustments for daily data are calculated on a monthly base and applied on a daily resolution, depending on which quantile of the monthly distribution the daily data belong to. The distribution of temperatures is then considered for each month separately, introducing the seasonal cycle in the adjustments. Absolute temperatures from the target month m are considered, and values from the preceding month and the following month are gathered to reduce the noise and to make sure that a sufficient amount of data are available to determine the quantiles. This approach makes that weather types from the spring and autumn transition seasons, like March and May, contribute to determining the adjustment for homogenization of April temperatures, which is likely to be influenced both by typical March and May conditions. These temperature measurements are sorted in ascending order and e.g. the value associated to 10^{th} quantile is calculated as the median value of all data points between the 7.5^{th} and 12.5^{th} quantiles. This process generates quantile sequences for data before and after the break in the target ($\mathbf{s}_{q,m}$,$\mathbf{b}_{q,m}$) and in the reference series ($\mathbf{r}^{\text{bef}}_{j,q,m}$,$\mathbf{r}^{\text{aft}}_{j,q,m}$) (thus obtaining 4 quantile sequences).

The adjustment for each of the quantiles is then calculated in a three-step approach. First the difference between quantile sequences of the candidate series before and after the breaks is calculated, this difference is affected by both the artificial and the climatic signal. Secondly, to identify the climatic signal, the difference between quantile sequences of the reference series before and after the break is calculated. As third step, in order to isolate the artificial signal, the above differences are subtracted to each other. These steps are summarized by the following equation:

$$\mathbf{a}_{i,j,q,m} = (\mathbf{b}_{q,m} - \mathbf{s}_{i,q,m}) - \left(\mathbf{r}^{\text{aft}}_{j,q,m} - \mathbf{r}^{\text{bef}}_{j,q,m}\right) \tag{2.2}$$

In order to reduce noise, the adjustments are smoothed using a simple mean of adjustments from the neighbouring months and neighbouring quantiles.

$$\bar{\mathbf{a}}_{j,q,m} = \frac{\mathbf{a}_{j,q,m} + \mathbf{a}_{j,q+5,m} + \mathbf{a}_{j,q-5,m} + \mathbf{a}_{j,q,m+1} + \mathbf{a}_{j,q,m-1}}{5} \tag{2.3}$$

Further check has been performed to avoid situations in which negative slopes of the smoothed sequences cause, after their application, changes in the rank of the data (i.e. data changing quantile after homogenization). This check has interested a very small portion of the series, more details on this can be found in Appendix 2.A

As mentioned above, the application of adjustments is performed considering each daily value of the series individually, depending on the location in the monthly temperature distribution.

A set of estimations of the correction is produced, each one corresponding to the different overlapping periods each reference series $\mathbf{R}_j$ has with the segments of the candidate series. The value to be corrected may belong to a different quantile in each of these overlapping periods. After determining these quantiles ($\tilde{q}_j$) the estimation ($\tilde{v}_j$) of the adjusted value related to $\mathbf{R}_j$ is:

$$\tilde{v}_j = v + \mathbf{a}_{j,\tilde{q},m} \tag{2.4}$$

where v is the original value. The final adjusted value is then calculated taking the median of the estimations:

$$\overline{v} = \text{median}(\tilde{v}_j) \tag{2.5}$$

where j=(1,...,r).

The method described above has been applied to the whole ECA&D dataset. The high number of breaks detected caused a high number of short homogeneous segments, which often were not long enough to be homogenized or had too short overlapping period with the homogeneous segments of the surrounding reference series, making it impossible to perform the quantile matching. These portions have been integrated in the temporary homogenized version of the dataset, which has undergone a second round of homogenization. In this second run the break detection and quantile matching have been launched again, so that additional adjustments were calculated.

2.3 Results

2.3.1 Statistics of the adjustments

Figure 2.3 shows the number and timing of the detected breaks for the first iteration and for the second iteration. The relatively high number of breaks detected in the original series during the first iteration resulted in a high number of short homogeneous segments. These segments, serving as references, are often not long enough to be homogenized or had a too short overlapping period with the target series. In such a situation no adjustments are possible. Nevertheless, the first homogenization iteration improved the number and the length of the homogeneous segments which made it possible to adjust additional breaks in the second iteration. The number of series for daily maximum and minimum series in the ECA&D dataset and the number of adjusted series after the first and second iterations are shown in Table 3.1.

Since the second iteration takes the results of the previous iteration as input, the possibility exists that this complex system diverges too strongly from the initial situation (e.g. positive feedbacks in the iterative processes, tendency for the removal of local signals or introduction of forced trends in the temperature series). This issue might be alleviated by applying the homogenization on the original series, using the homogenized series as references in each iteration and setting a convergence threshold to stop the system. In the approach documented here, we simply limit the number of iterations to two.

Table 2.1: *Number of series involved in the stages of the homogenization process. Homogenized series, for each iteration, consist of the sum of the adjusted series and the already homogeneous series. Low percentage of homogenized series is due to the exclusion of short and duplicate series.*

	TN	TX
Original series (complete set of considered series)	6438	6404
Homogeneous original series (original series labelled as homogeneous by the break detection)	560	670
Adjusted series it.1 (series that have been corrected during iteration 1)	2111	2007
Homogenized data-set, iteration 1 (union of series that were already homogeneous and adjusted series)	2671	2677
Homogeneous series after iteration 1 (union of original homogeneous series and series successfully homogenized after iteration 1)	1165	1131
Adjusted series, Iteration 2 (series that have been corrected during it.2)	1526	1571
Homogenized data-set (final) (union of series that were already homogeneous and adjusted series)	2691	2702
Final non homogeneous series (according to a third run of break detection)	1357	1260
Final homogeneous series (according to a third run of break detection)	1334	1441

The adjustments that have been applied to the breaks vary strongly with the month and the quantiles. Figure 2.4 shows that the adjustment of the median is symmetric around -0.1°Cand smaller (and less frequent) adjustments are found for the second iteration. Furthermore the distribution of adjustments of the median is wider for minimum temperatures and more peaked around for maximum temperature. Averages of adjustments for the 5th , 50th and 90th quantiles (see table 2.2) are more negative for TN. For both variables adjustments for lower quantiles are more negative, indicating a general tendency

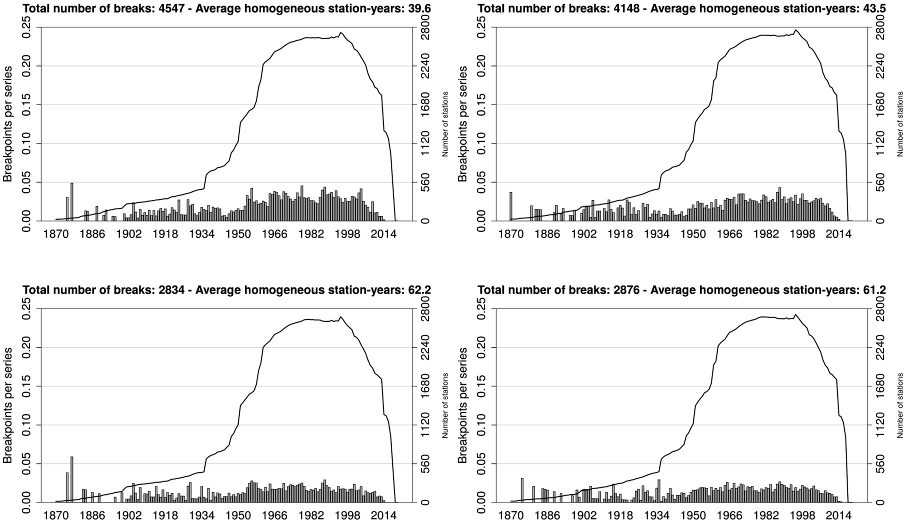


Figure 2.3: *Statistics regarding application of break detection on ECA&D temperature dataset. Histogram describes number of breakpoints per series, line describes number of stations. Left (Right) panels are about minimum (maximum) temperatures. Top and bottom panels are respectively for first and second break detection runs.*

to broaden the probability density distribution. This is consistent with earlier findings [Lawrimore et al., 2011; Trewin, 2013; Thorne et al., 2016].

Table 2.2: *Averages of adjustments for 5th , 50th and 95th percentile for TN and TX in first and second iteration*

	TN			TX		
	05th perc	50th perc	95th perc	05th perc	50th perc	95th perc
1st it.	-0.12 °C	-0.11 °C	-0.10 °C	-0.06 °C	-0.05 °C	-0.06 °C
2nd it.	-0.03 °C	-0.01 °C	-0.01 °C	-0.04 °C	-0.03 °C	-0.03 °C

The peak in the distribution close to zero relates to (a) the independence between break detection and adjustment calculation, i.e. a break may be found but the comparison to the surrounding reference series does not give a reliable correction and (b) to the possibility that the median needs no adjustment but percentiles in the tails do.

This latter situation is illustrated in the scatterplots of adjustments for the 5th versus the 95th quantiles, for the series where adjustments of the median are null (figure 2.5). These figures show a centred and symmetric distribution and indicate that adjustments are not skewed towards more positive or negative slopes. These figures also show that no thresholds on the absolute value of the adjustments are used, contrasting with the approach of [Trewin, 2013] who applied adjustments only if the resultant shift in annual mean exceeded the ± 0.3 °C threshold.

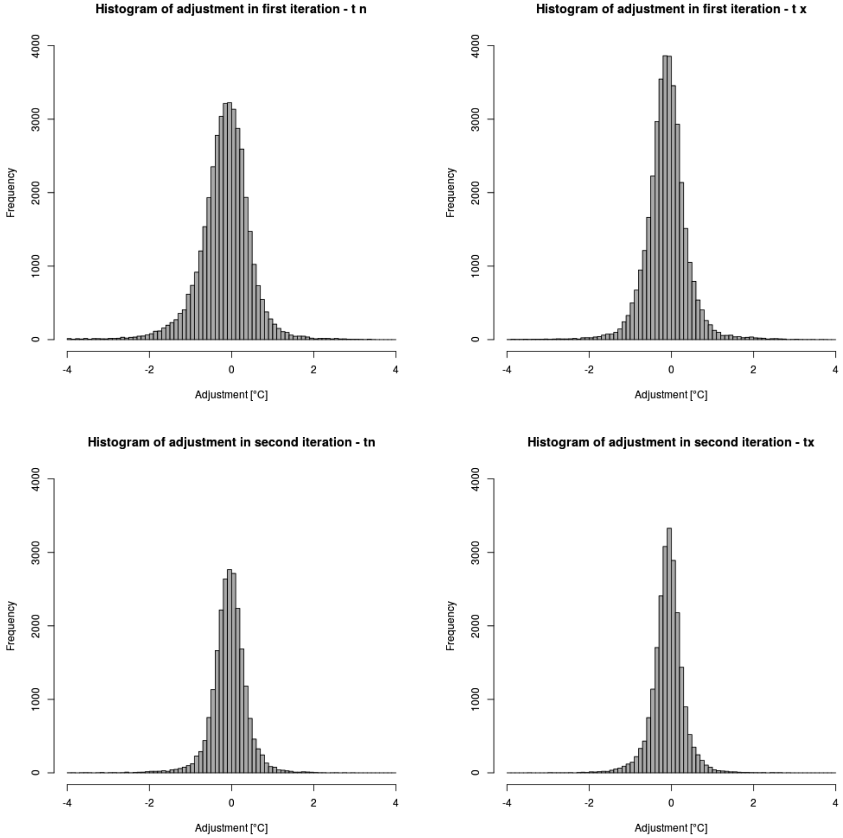


Figure 2.4: *Histograms of adjustments for values in the median quantile (q50), for TN (left column) and TX (right colums) and for first iteration (top row) and second iteration (bottom row). Difference in width between first and second iteration proves the different role of the two phases.*

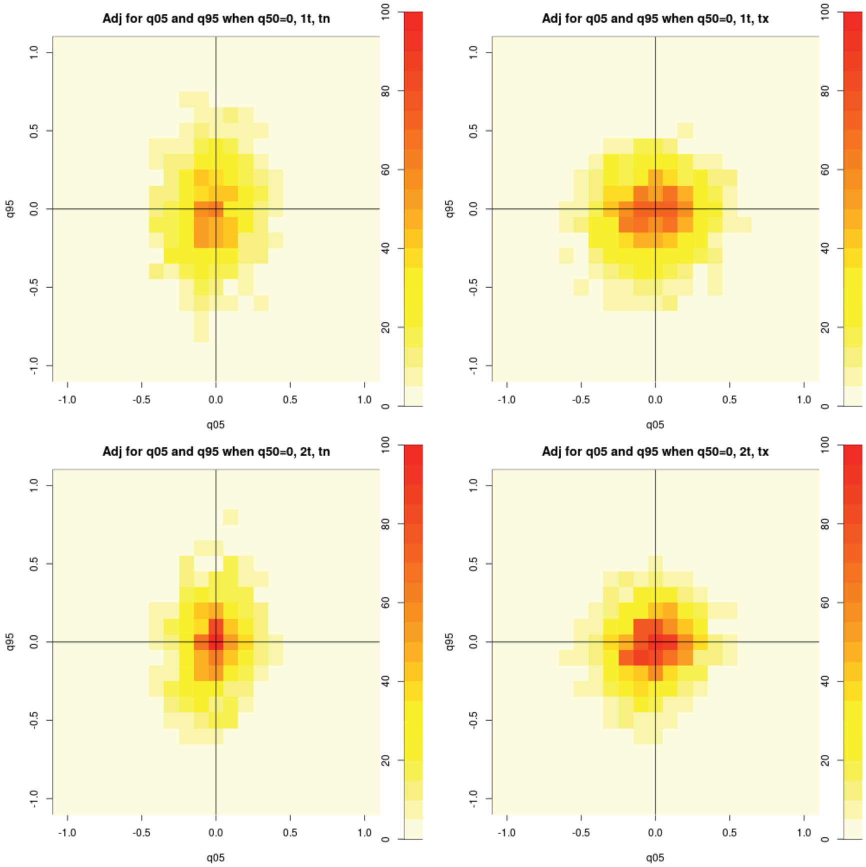


Figure 2.5: *Density scatterplot of adjustments for quantile 95 versus quantile 05 when adjustments for the median are null. Minimum temperature (left column), maximum temperature (right column), first iteration (top row), second iteration (bottom row).*

2.3.2 Case studies

In order to demonstrate the method in more detail, two case studies are presented.

Bamberg

An illustrative example is the adjustment of data from the station of Bamberg (Germany) [1]. Metadata reports a set of breaks (Table 2.3) which are only partially retrieved by the automatic break detection (first iteration: 1891,1952; second iteration: 1920).

The two documented breaks are not reproduced exactly, but the 1948/49 break is located within a few years. The further breaks that are detected are probably related to unrecorded changes in the features of the station.

Table 2.3: *Available metadata regarding station in Bamberg, Germany.*

1880-01 to 1948-12	Bamberg (Sternwarte - City Centre)	z=283m
1949-01 to 1995-03	Bamberg (South - Country Side)	z=243m
1995-03 to present	Bamberg (South - Country Side)	z=239m

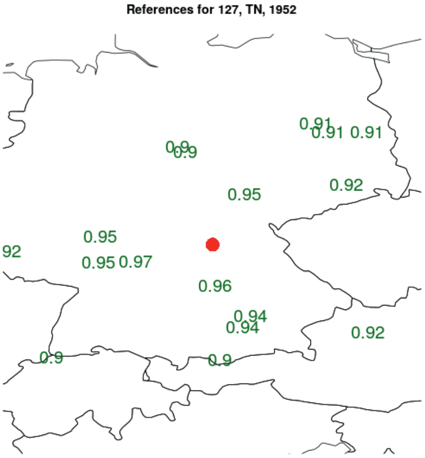


Figure 2.6: *Reference series that have been used for the homogenization of the break in 1952 in the station of Bamberg (red dot). Numbers represent correlation calculated between references and basis series (considering the 20 years following the break).*

[1] The German Weather Service now makes the data for station Bamberg available as separate series ranging from 1879 to 1958 and from 1949 to 2018

The high density of stations in Germany and Austria provided by their respective meteorological services allows to have more than 18 reference series available for the break in 1952 on which we focus in this case study. The 18 highest correlated ones have been selected (figure 2.6).

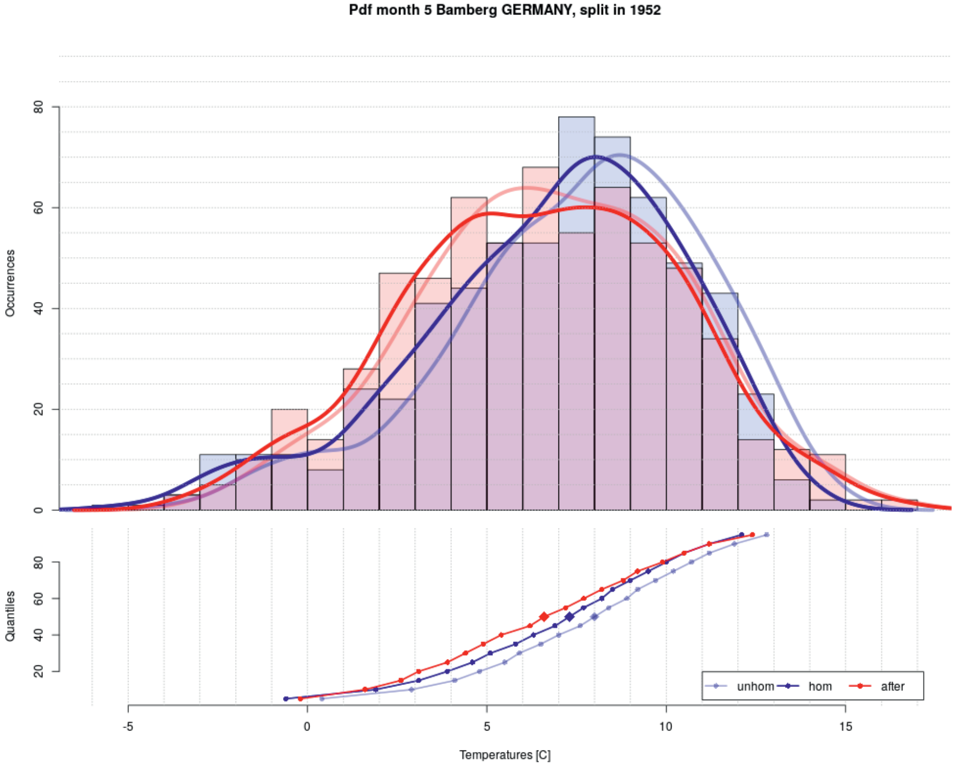


Figure 2.7: *Top: Histograms and pdf of adjusted minimum temperature in month May for the 20 years before 1952 (blue) and 20 years after that (red). Light blue and light red curves represent original probability density functions. Bottom: Quantile functions related to the above distributions, same colour code.*

Shape and location of the probability distributions of the non-homogenized temperatures before (light red) and after (light blue) the break show a clear distinction (figure 2.7). Shifts in quantile sequences varies from very low values for the tails of the distribution to 1.1°Cfor the median, showing the different effect of the break on mean and extreme values. Probability plots after the two iterations of homogenization (red before and blue after the break) get closer to each other in different way depending on the quantiles, indicating that the difference between the two original distributions was not entirely due to the artificial intervention. The two sub-series (before and after the break) do not completely overlap due to the climatic variability that has been captured by the surrounding reference series and taken into account in determining the adjustments.

Estimates of adjustments related to each reference (for this case study) are shown in figure 2.8. These are the results of the process described by equation 3.1, followed by smoothing and check of negative slopes. Figure 2.8 shows that the lower quantiles have stronger (more negative) adjustments than the upper quantiles which increases the width of the distribution. As an example of the adjustment process, the estimates related to the value measured on 22th May 1951 (6.2°C) are highlighted in figure 2.8. This measurement belongs to different quantiles (35th and 40th), depending on the reference series that is considered, since for each of these there is a different overlapping period with the target series. The final correction is taken as the median of the estimates, in this particular case the adjustment will be -1.3°C, with a final homogenized value of 4.9°C.

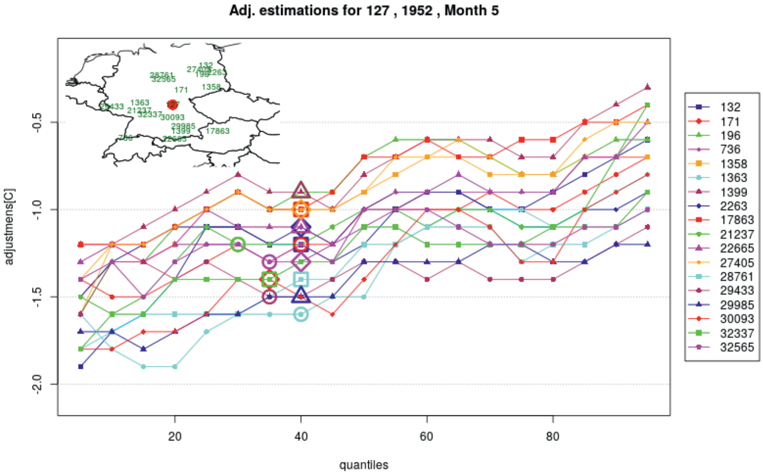


Figure 2.8: *Estimation of adjustments for month May , station of Bamberg and break in 1952 after the smoothing process and the negative slop check.*

Effects on the series are evident when indices like the annual mean are plotted (figure 2.9, top panel). In this particular case, the first iteration is able to correct the series almost entirely, since the break that has been detected during the second iteration (1920) had very low adjustments. Comparison of corrections for the mean and for the two tails of the distribution show the expected differences: larger(smaller) corrections for the 5th (95th) quantile.

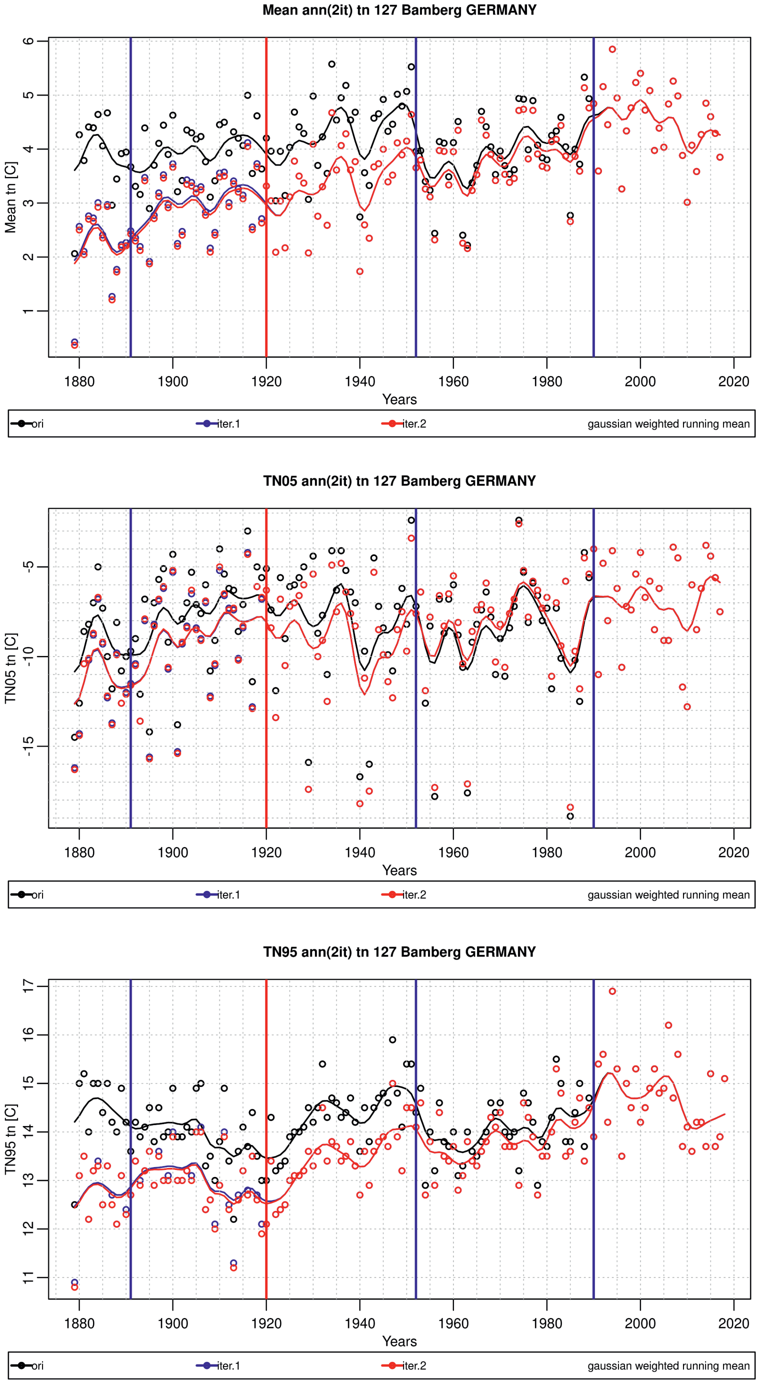


Figure 2.9: *Annual mean (top), 5th quantile (centre), 95th quantile (bottom) time series for minimum temperatures in Bamberg. Black line: original series, blue: first iteration result, red: second iteration result. Vertical lines: output of break detection in the first (blue) and in the second (red) iteration.*

Salzburg

A particularly representative case is the station of Salzburg (Austria), where the metadata reports a set of breaks (table 2.4). Break detection detects almost all these breaks, except the most recent one which is probably small in amplitude (only 3 meters of change in height), and detecting some further breaks which probably derive from unreported changes in the stations features. For this case study, focus will be on the break located near 1938 which is associated with the relocation of the station from the city to the nearby airstrip.

Table 2.4: *Available metadata for station in Salzburg, Austria.*

1863-01 to 1883-12	High School (Gymnasium Altes Borromäum)	z=424m
1884-01 to 1903-07	High School (Oberrealschule)	z=419m
1903-08 to 1941-02-28	Studiengebäude-Lehrerbildungsanstalt	z=423m
1939-03-01 to 1996-06-15	Airport station 1	z=434m
Since 1996-06	Airport station 2	z=437m

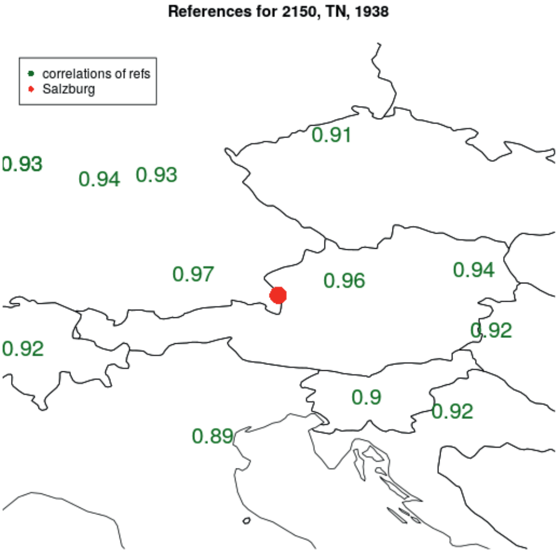


Figure 2.10: *Reference series that have been used for the homogenization of the break in 1938 in the station of Salzburg (red dot). Numbers represent correlation calculated between references and basis series. Correlations are calculated using the 20 years following the break.*

For the 1938 break, 12 reference series meet the requirements, see figure 2.10 . Shape and location of the probability distribution of temperatures before and after break in 1938 (figure 2.11) show a clear shift of the distribution of records before the break. Shift in quantiles varies from 0.6°Cfor the 5th quantile to 1.1°Cfor the median, showing the different effect of the break on mean and extreme values. Probability distributions after the 2 iterations of homogenization almost overlap each other, indicating that a great part of the difference was due to the relocation, while the remaining difference represents actual climatic variability that relates to the surrounding reference series.

Adjustments applied may be seen in figure2.12, analogous to figure 2.8 with the difference that here the slope of the curve in the quantile-adjustment plot is not as steep as in the Bamberg case. The highlighted marks are related to the measurement on 14th May 1938 (4.9°C), whose adjustment will be -0.7°C, with a final value of 4.2°C.

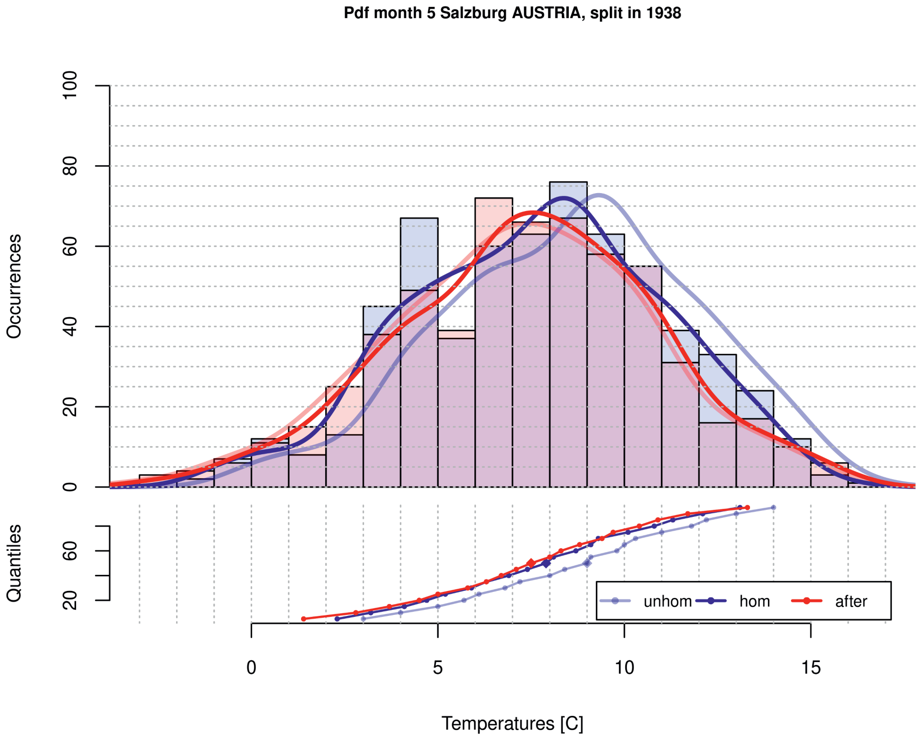


Figure 2.11: *Top:Histograms and probability distribution of adjusted minimum temperature in month May for the 20 years before 1938 (blue) and 20 years after that (red). Light blue and light red curves represent original distributions. Bottom: quantile functions related to the above distributions, same colour code.*

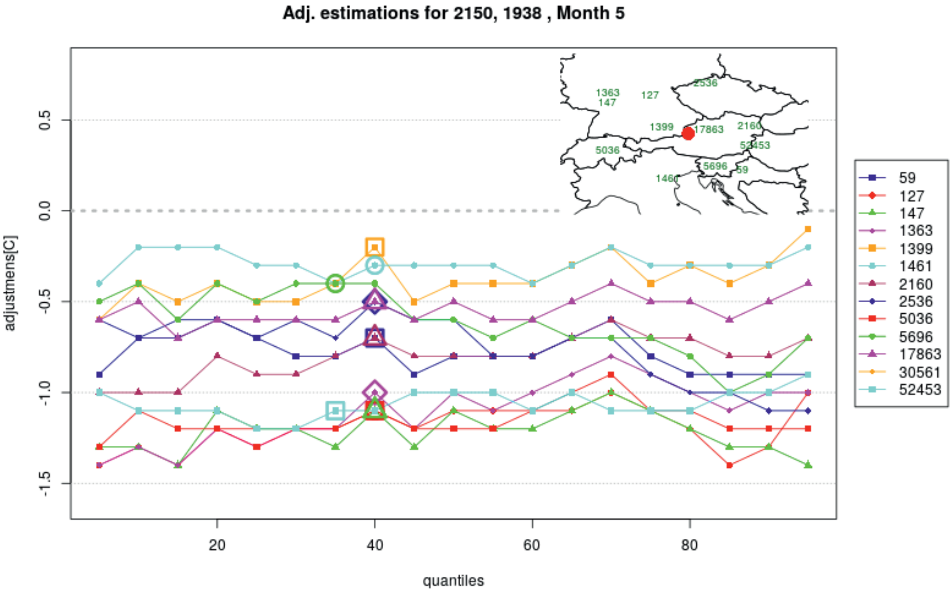


Figure 2.12: *Estimation of adjustments for month May , station of Salzburg and break in 1938 after the smoothing process and the negative slop check. The inset shows the locations of the series used to calculate the adjustment.*

Effects on the series are evident when indices like the annual mean are plotted (figure 2.13, top panel). Interesting about this case is that first iteration (blue lines, when not covered by red) corrects the big breaks, such as the break in 1938. On the other hand, the second iteration is able to adjust two early breaks (red vertical lines) that were not detected during the first round because of the lack of long reference series in the early periods. The two new breaks are confirmed by the metadata (table 2.4). The amplitude of adjustments in this case is clearly lower, showing that second iteration works as an enhancement of adjustments from the first round. Appendix 2.B shows that the second iteration homogenizes the older part of the series that was not corrected during the first round.

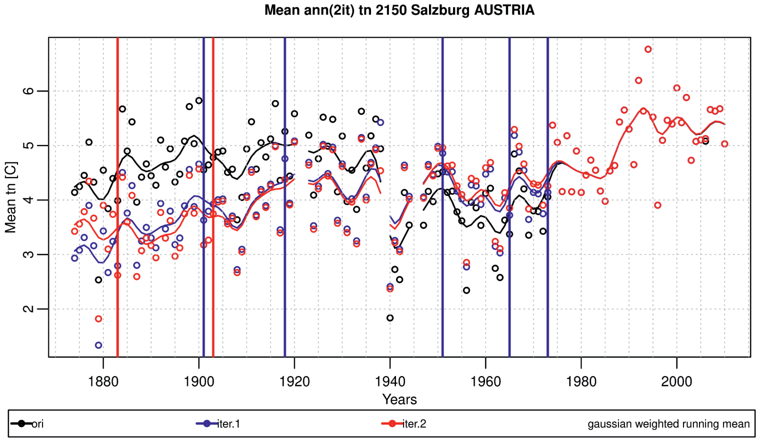


Figure 2.13: *Annual mean time series for minimum temperatures in Salzburg. Black line: original series, blue: first iteration result, red: second iteration result. Vertical lines: output of break detection in the first (blue) and in the second (red) iteration.*

2.3.3 Application to the complete data-set

Figure 2.14 shows trends in annual mean daily minimum and maximum temperature over the 1961-2010 period, before the homogenization (top panels) and after the homogenization (middle panels). The comparison of these figures shows the removal of several outliers and unrealistic low or high trends values. The trends based on the homogenized series are much more spatially homogeneous. The bottom panels of figure 2.14 show the difference in trend values between the non-homogenized and homogenized series which demonstrate that the adjustments go both ways - trends are increased and decreased by this procedure. A coherent spatial pattern of adjustments is not evident from this figure.

The almost complete disappearance of stations with negative and very large trends demonstrates the effectiveness of the method in recognizing and keeping the climate signals that dominates the series and removing outliers trends which are related to artificial signals. Even though the result is the convergence of the trends of all stations to positive values, it is important to notice that the aim of this process is not the removal of the negative trends. This phenomenon is an indirect effect of the homogenization procedure. Indeed all stations with excessively high trends (i.e. dark red circles) have been adjusted with negative factors, as shown in figure 2.14, bottom row. A further check is to search for stations that still showed a negative trend or a exceptionally high trend exceeding 0.6°C/dec. Figure 2.15 shows the locations of the stations related to these extreme trends. These are not isolated stations but it is shown that these values are consistent with trends of neighbouring stations. The low value trends are mainly located in Bulgaria and southern

Romania, while the very large trends are mainly in the Northern Baltic area. The second case is likely to be the result of a widespread climatic effect, while the first might be the result of the influence of the series of Bucarest on the neighbours.

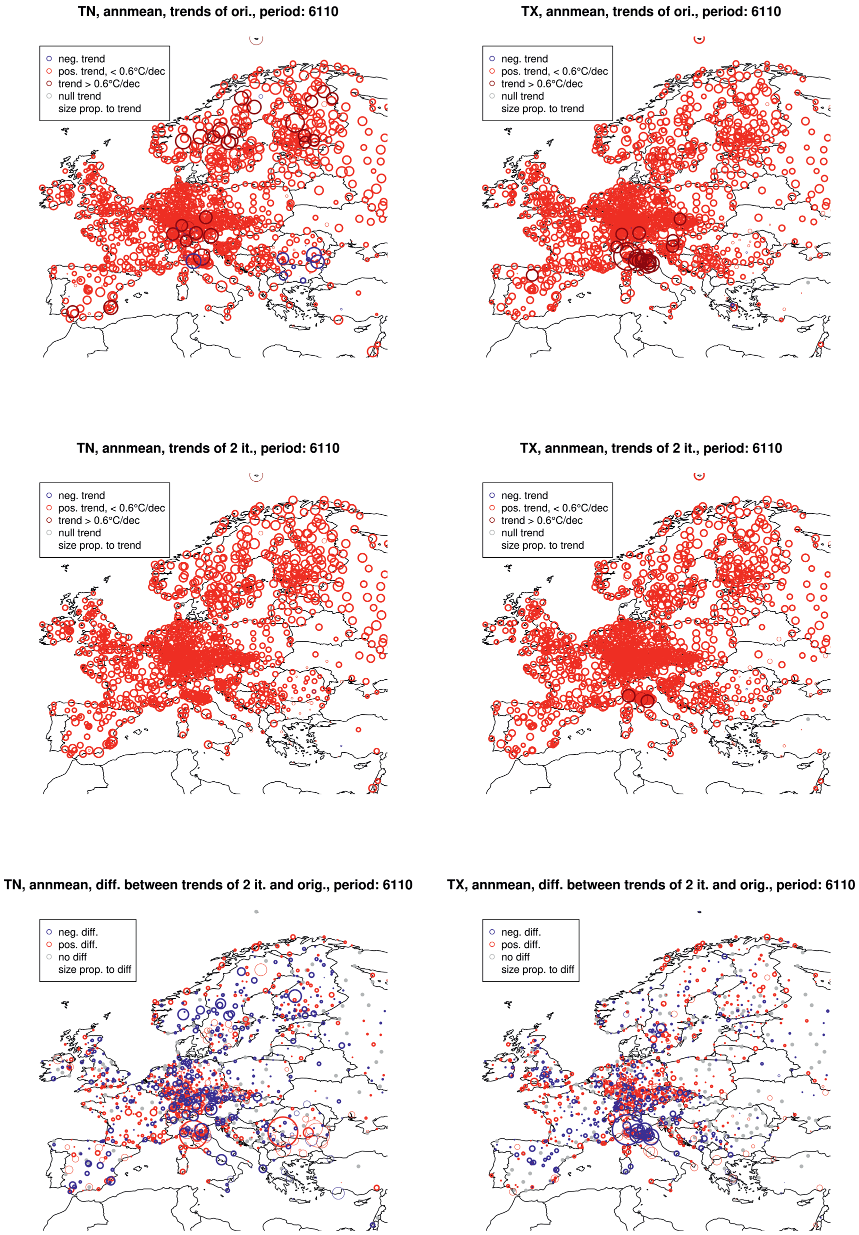
In Appendix 2.C box plots show the distribution of the trends in the annual mean of TN and TX for the two successive iterations. These indicate a narrowing of the distribution of the trends together with a significant reduction of outliers.

Figure 2.16 describes the distribution of the trends on extreme indices (5^{th} percentile of TN and 95th percentile of TX) in the original data-set and after the two stages of homogenization.

Beside the changes in the width of the distributions, changes in the first moment have been observed. The medians show a slight shift to higher values (table 2.5) for annual means. TN05 and especially TX95 show a more irregular behaviour, where overadjusted trends in the first iteration are refined in the second iteration.

Table 2.5: *Median trend values of annual averaged daily minimum and maximum temperature, of TN05 and of TX95, for the original dataset, after the first iteration and after the second iteration.*

	TN annmean	TX annmean	TN05	TX95
original	+0.31 °C/dec	+0.35 °C/dec	+0.41 °C/dec	+0.35 °C/dec
1^{st} iteration	+0.31 °C/dec	+0.37 °C/dec	+0.42 °C/dec	+0.40 °C/dec
2^{nd} iteration	+0.32 °C/dec	+0.37 °C/dec	+0.42 °C/dec	+0.37 °C/dec



TN, annmean, trends of ori., period: 6110

TX, annmean, trends of ori., period: 6110

TN, annmean, trends of 2 it., period: 6110

TX, annmean, trends of 2 it., period: 6110

TN, annmean, diff. between trends of 2 it. and orig., period: 6110

TX, annmean, diff. between trends of 2 it. and orig., period: 6110

Figure 2.14: *Maps of trends of annual mean in the period from 1961 to 2010 of original series (top), homogenized series (middle) and difference between the two (bottom) about minimum (left column) and maximum temperatures (right column). Blue circles indicate negative trends, red circles represent positive trends below 0.6° C/dec, dark red circles represent trends above 0.6° C/dec. Size of the circle is proportional to the amplitude of the trend. Thickness of the circle indicates significance of the trend itself (above 0.95). Code colour is chosen basing on the box plots of figure 2.18 (blue and brown values lie on the tails of the original distribution).*

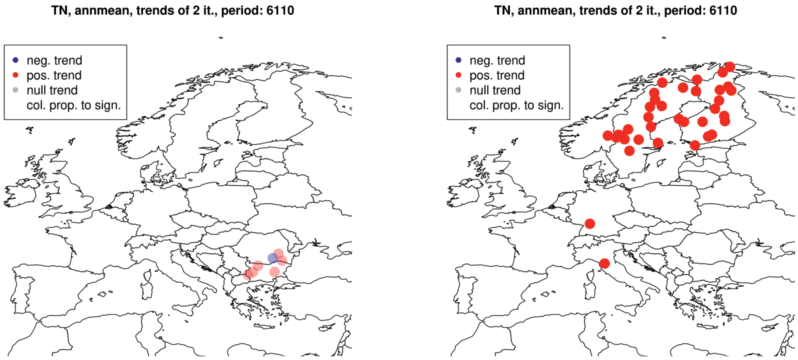


Figure 2.15: *Left: Map of series having trend of annual mean of minimum temperatures lower than 0.1° C/dec, transparency of the dots indicate non significant trend. Right: Map of series having trend of annual mean of minimum temperatures larger than 0.5° C/dec, transparency of the dots indicate non significant trend.*

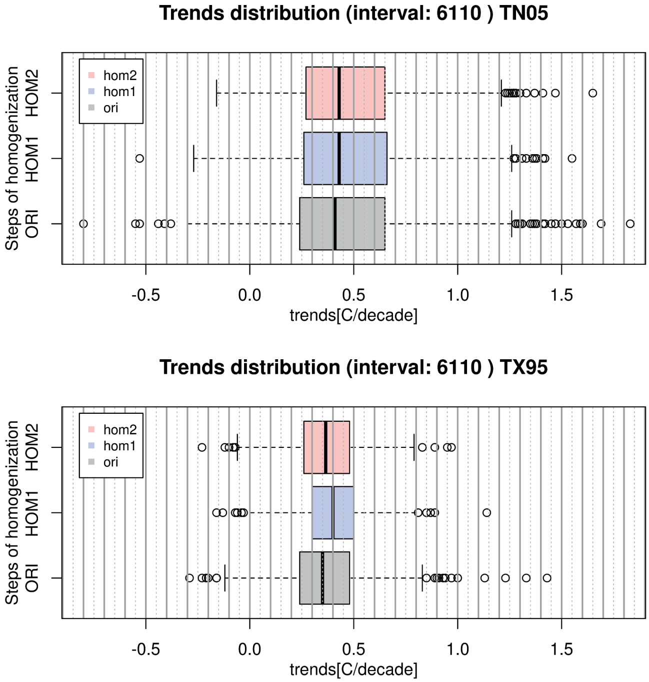


Figure 2.16: *Distribution of trends for TN05 (top) and TX95 (bottom). Each plot shows boxplot for original (grey), first iterations' result (blue) and second iterations' result (red)*

2.4 Discussion and Conclusions

A fully automatic homogenization method for daily temperature series has been presented and applied to the pan-European dataset ECA&D. The size of the dataset and the absence of detailed metadata for all series and stations requires a procedure which is *blind* for metadata and is able to handle a large variety of data quality conditions. These challenges are met by using a combination of the break detection method developed earlier by [Kuglitsch et al., 2012] and the quantile matching method which has been pioneered earlier by [Trewin, 2013] for use in large datasets. In order to distinguish between climatic signal and artificial signal in the breaks, a network of reference series in the vicinity of the record that needs adjustment is employed. In this study, the reference series are chosen from coordinate boxes of $6° \times 6°$ around the target series, with thresholds on altitude difference. A further selection, using daily raw correlation and length of overlapping period, makes this approach sufficiently flexible to cope with both high and low station density areas. These criteria take inspiration from the the nearest neighbour stations approach of [Menne and Williams Jr, 2009] and [Trewin, 2013].

The whole procedure is iterated twice. While the first iteration locates and adjusts the largest discontinuities, the second iteration is able to adjust more subtle changes such as earlier breaks, breaks with smaller amplitude or in areas with scarce station density. The more abundant presence of homogeneous sub series after the first iteration makes this possible. The homogenization has been able to adjust about 2700 ECA&D temperature series, while the remaining ones are duplicates or consist in short records. A final round of break detection has shown that only 1400 series can be considered completely homogenized, while on the rest minor breaks persist. No further iterations are made to remain as close as possible to the original dataset while adjusting for the largest inhomogeneities.

The trends in annual averaged values show a much stronger spatial consistency than before the adjustments. This illustrates the effectiveness in the removal of the artificial signals, thus making the climatic signal dominant. A comparison between trends prior and after the homogenization shows that changes in trends are both ways. The averages of the distributions of European trends of annual means are shifted slightly to warmer values (TN: +3.2%, TX: +5.7%). At the same time the interquantile range of the distributions of these trends are consistently reduced (TN: from 0.16 to 0.10°C/dec, TX: from 0.12 to 0.08°C/dec), indicating a higher uniformity of the values. Similar conclusions are reached when considering indices for extreme values, such as the 5th and 95th quantiles.

The strength of the quantile matching method is that each part of the distribution (i.e. quantile) is considered independently from the others. Previous studies on temperature probability distributions have focused on fitting the probability density functions with sophisticated functions or calculation of variations in the distribution parameters. On the other hand the quantile matching approach has a more heuristic approach, aiming at being more versatile and able to adapt to the wide spectrum of signals that artificial activities may lead to the records.

In distinction with the Trewin [2013] approach, no linear interpolation between quantiles (used to obtain adjustments for percentiles between the multiples of 5) is included, reducing the parameterization of the process. Furthermore the calculation of adjustment is more conservative (use of averaged values and check of negative slopes) and the selection of reference series employs different criteria to give more importance to data availability and correlation.

Nonetheless further studies have been performed and are planned to understand minor controversial aspects of the described method. The dispersion of the reference series has been shown to affect the calculation of the adjustments (Appendix 2.D). Therefore it is planned to inspect how to lend the reference selection an ”angular even distribution”, i.e. approximately same number of references on the North, on the South, on the East and on the West of the candidate series. The selection of reference series must also take into account the contribution of series with anomalous behaviour, as seen in Figure 2.15, where the negative trend in a station (Bucarest) might be one of the reasons of the lower trends observed in the surrounding area. The validation and the comparison of the results with other adjustment calculation methods are currently subject of further studies.

In conclusion, the method to adjust inhomogeneities in daily temperature discussed here is a purely statistical method. While the use of quantile matching favours the differentiation of adjustments for low and high daily extremes and have a seasonal cycle, these adjustments don’t consider existing meteorological circumstances. Future work using a physical approach to calculating the adjustments, in which actual weather contributes to the size of the adjustment, will give an alternative estimate of the homogeneity adjustment.

2.5 Acknowledgements

We acknowledge the data providers in the ECA&D project (www.ecad.eu) and the helpful contribution of Finn Lindgren (University of Edinburg) in developing the mathematical bases for the development of the check of negative slopes (Appendix 2.A). Funding has been received from the EU FP7 Collaborative Project UERRA (Uncertainties in Ensembles of Regional ReAnanalysis), Grant agreement 607193, and the EU H2020 EUSTACE Project, Grant agreement 640171.

2.A Check of negative slopes in the adjustment sequences

During the application of the quantile matching method it might happen that the rank of measurements is not preserved. This occurs if the adjustment of a high quantile is smaller than that of a lower quantile.

This possible setback, that involves approximately 0.5% of the adjustment calculation, requires a constraint in order to keep the rank of data when the sequence has a negative slope in the adjustment - quantile plane. By definition a quantile sequence, including the result of the adjusting process ($\tilde{\mathbf{s}}_{j,q,m}$) must have a non-negative slope. This implies that for any q:

$$\bar{\mathbf{s}}_{j,q+5,m} - \bar{\mathbf{s}}_{j,q,m} \geq 0 \tag{2.6}$$

For each q, elements of the adjusted quantile sequence are calculated like:

$$\bar{\mathbf{s}}_{j,q,m} = \mathbf{s}_{j,q,m} + \mathbf{a}_{j,q,m} \tag{2.7}$$

Thus:

$$\begin{aligned} \tilde{\mathbf{s}}_{j,m,q+5} - \tilde{\mathbf{s}}_{j,m,q} &= \\ = (\mathbf{s}_{j,q+5,m} + \bar{\mathbf{a}}_{j,q+5,m}) - (\mathbf{s}_{j,q,m} + \bar{\mathbf{a}}_{j,q,m}) &= \\ = (\mathbf{s}_{j,q+5,m} - \mathbf{s}_{j,q,m}) + (\bar{\mathbf{a}}_{j,q+5,m} - \bar{\mathbf{a}}_{j,q,m}) &\geq 0 \end{aligned}$$

And finally:

$$(\bar{\mathbf{a}}_{j,q+5,m} - \bar{\mathbf{a}}_{j,q,m}) \geq -(\mathbf{s}_{j,q+5,m} - \mathbf{s}_{j,q,m}) \tag{2.8}$$

This constraint is implemented fixing the adjustment related to the median and checking the two tails quantile by quantile. In case of a too negative slope, the value is corrected moving it to the closest acceptable value.

For instance, for the right tail of he distribution, if:

$$\mathbf{a}(\tilde{j}, 55, \tilde{m}) - \mathbf{a}(\tilde{j}, 50, \tilde{m}) < -\mathbf{s}(\tilde{j}, 55, \tilde{m}) + \mathbf{s}(\tilde{j}, 50, \tilde{m}) \tag{2.9}$$

then the corrected value is set to:

$$\bar{\mathbf{a}}(\tilde{j}, 55, \tilde{m}) = \mathbf{a}(\tilde{j}, 50, \tilde{m}) - \mathbf{s}(\tilde{j}, 55, \tilde{m}) + \mathbf{s}(\tilde{j}, 50, \tilde{m}) \tag{2.10}$$

2.B Munich

The need and the utility of the second iteration with the break detection and homogenization can be appreciated when looking data from the station of Munich (Germany). Metadata reports a set of breaks (Table 2.6) including one in the 1920s.

Table 2.6: *Available metadata regarding station in Munich, Germany.*

1879-01 to 1954-07	Munich (Botanic Garden Nynphenburg)	z=515
1954-08 to 1999-03	Munich (Nynphenburg residential area)	z=515m
1999-04 to present	No location metadata (probably not changed) (changes in measuring times in 2001-04)	z=515m

Here the big difference between red and blue lines in figure 2.17 indicates that the first iteration was not able to correct the breaks in 1912 an 1926. Nevertheless the higher availability of data and long homogeneous segments, together with a better signal to noise ration has allowed to adjust the earliest part of the series on the second run of the software.

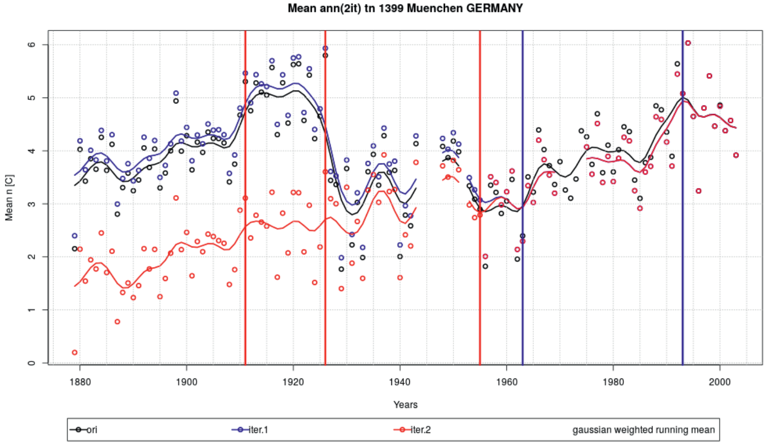


Figure 2.17: *Annual mean time series for minimum temperatures in Munich. Same colour code as previous figures.*

2.C Trend assessments on annual means

Assessment of trends before and after the homogenization has been computed on annual means, showing a relevant narrowing of the distribution, especially between original and first iteration, while the second iteration acts more like a refining of the result, see figure 2.18.

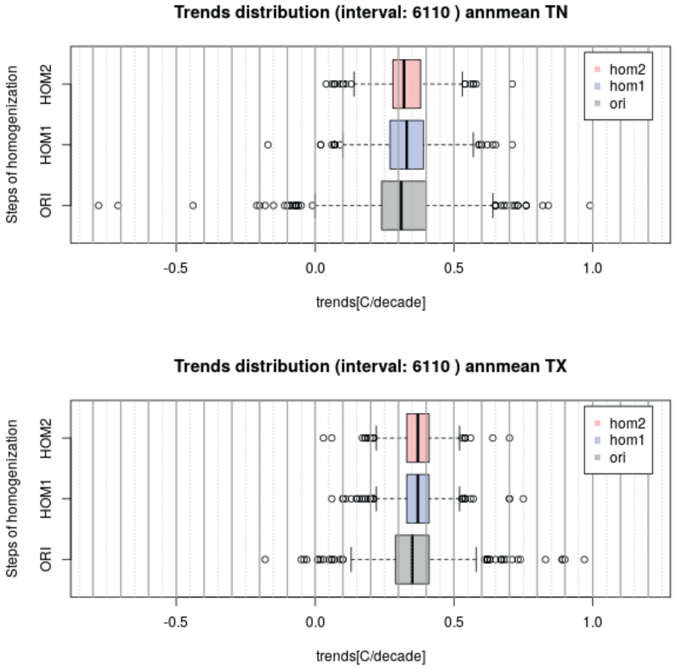


Figure 2.18: *Distribution of trends for annual mean of minimum (top) and maximum temperature. Each plot shows boxplot for original (grey), first iterations' result (blue) and second iterations' result (red)*

2.D Geographically-induced patterns on adjustments

Adjustment calculations may depend on the geographical distribution of reference series around the target series. The homogenization of the break in 1938 of series of Salzburg (Austria) has been checked considering separately the reference series that lay on the north (south, west and east), see figure 2.19, top panels. Both pairs north-south and east-west present some differences, more evident in the case east-west. Reason of this difference is highlighted in the adjustments sequences for month of May of western and eastern series, where it is clear that four series in the western dataset introduce larger (negative) adjustments. This series correspond to the 4 north-western German stations, which differ for topographical characteristics with respect to the rest of the stations.

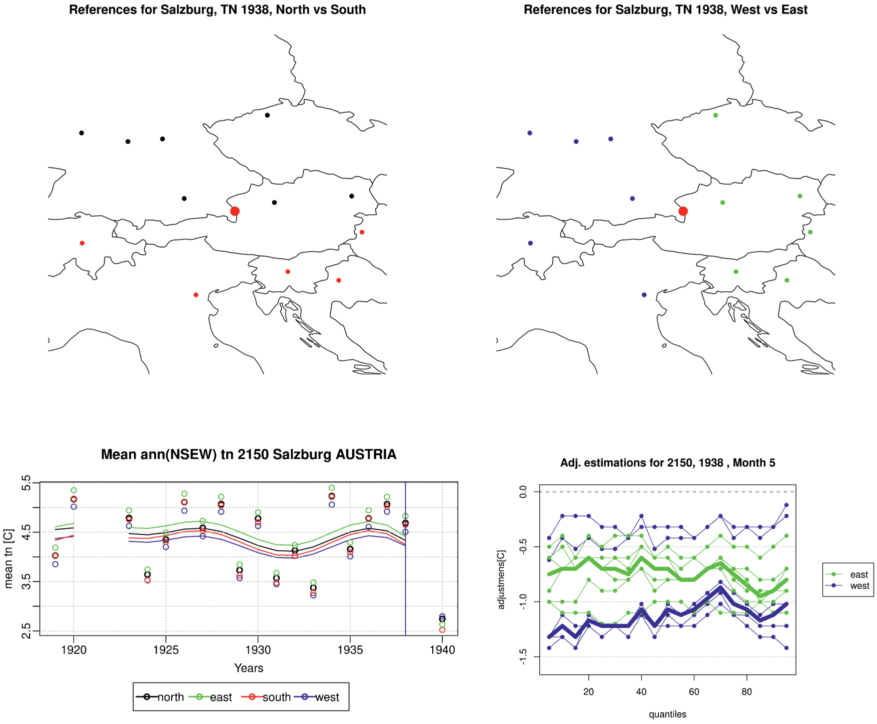


Figure 2.19: *Annual mean of minimum temperatures in Salzburg after homogenization using 4 different sets of reference series: northern (black), eastern (blue), southern (red), western set (green).*

Therefore the extremely high variety of European topography and climate features require to perform an accurate choice of the references for each target series, reasoning on its locations, surroundings etc. Furthermore in some cases the sparseness of stations density doesn't allow to be able to select the best stations. In further versions of the software an even angular distribution of the references around the target will be implemented in case of regions with high density of stations.

Chapter 3

Building long homogeneous temperature series across Europe: a new approach for the blending of neighboring series

Abstract

Long and homogeneous series are a necessary requirement for reliable climate analysis. Relocation of measuring equipment from one station to another, such as from the city center to a rural area or a nearby airport, is one of the causes of discontinuities in these long series which may affect trend estimates. An updated procedure for the composition of long series is here introduced. It couples an evolution of the blending procedure already implemented within the European Climate Assessment and Dataset (which combines data from stations no more than 12.5 km apart from each other) with a duplicate removal, alongside the quantile matching homogenization procedure. The ECA&D contains approximately 3000 homogenized series for each temperature variable prior to the blending procedure, around 820 of these are longer than 60 years; the process of blending increases the number of long series to more than 900. Three case studies illustrate the effects of the homogenization, showing the effectiveness of separate adjustments on extreme and mean values (Geneva), on cases where blending is complex (Rheinstetten) and on series which are completed by adding relevant portions of GTS synoptic data (Siauliai). Finally, a trend assessment reveals the removal of negative and very large trends, demonstrating a stronger spatial consistency. The new blended and homogenized data-set will allow a more reliable use of temperature series for indices calculation and for the calculation of gridded data-sets, and will be available for users on www.ecad.eu.

This chapter was originally published as:
Squintu, A. A., van der Schrier, G., van den Besselaar, E. J., Cornes, R. C., and Klein Tank, A. M. (2020c). Building long homogeneous temperature series across europe: a new approach for the blending of neighboring series. *Journal of Applied Meteorology and Climatology*, (2020)

https://doi.org/10.1175/JAMC-D-19-0033.1

3.1 Introduction

Long and high resolution temperature series are fundamental in climatological studies for giving a historical perspective of average warming and climatic extremes [Peterson et al., 1998; Aguilar et al., 2003]. An important requirement for these series is that they are homogeneous. The removal of non-climatic or artificial signals, *inhomogeneities*, is essential since climatic trends may be contaminated by these signals as shown by e.g. Venema et al. [2013] and Squintu et al. [2019].

It is highly likely that the surroundings of long-running meteorological stations will change over time or that the location of the instruments and possibly also the measurement procedure will also change [Domonkos, 2011; Kruger and Nxumalo, 2017; Vincent et al., 2018]. In addition, stations may be temporarily or permanently terminated, producing discontinuities in the dataset.

A common procedure [ECA&D Project Team, 2012; Menne et al., 2012; Kruger and Nxumalo, 2017] that is used to build long and continuous time series is to fill gaps in a series using data from neighboring stations or to combine long segments of measurements. When data of stations from very different surroundings are combined, in particular from urban and rural (e.g. airport) areas, large inhomogeneities are introduced [Trewin and Trevitt, 1996; Tuomenvirta, 2001; Brunet et al., 2006; Böhm et al., 2010; Rahimzadeh and Nassaji Zavareh, 2014; Vincent et al., 2018]. The extension of series can be also performed with data related to sources that are not completely comparable. An example of this is the update of series to recent periods with GTS-derived synoptic messages [Van den Besselaar et al., 2012], *synops*. The presence of these discontinuities makes the combined data unreliable for accurate climatological analyses, since the amplitude of inhomogeneities can be as large as the climate change signal itself [Peterson et al., 1998; Caussinus and Mestre, 2004; Begert et al., 2005; Della-Marta and Wanner, 2006; Venema et al., 2013].

Several studies on singular cases or on small datasets have produced tailored homogeneity adjustments for composite series [Maugeri et al., 2002; Böhm et al., 2010; Yang et al., 2013; Dienst et al., 2017; Kruger and Nxumalo, 2017; Delvaux et al., 2018; Nemec et al., 2013], and were able to account for the specific characteristics of the considered series and for metadata. In the case of large datasets, it is not possible to generalize a tailored approach for an individual case to the whole dataset due to the number of series involved [Dienst et al., 2017; Delvaux et al., 2018]. For such datasets a fully automatic procedure is required which has to be able to handle the large spectrum of, sometimes unknown, circumstances (different surroundings, measurement techniques, protocols and elevation) that each specific case presents. The procedure can not rely on metadata as this information is often lacking

The aim of this study is to provide a new approach to the operational procedure used in the construction of composite series (*blending*) [ECA&D Project Team, 2012] in the European Climate Assessment & Dataset (ECA&D) [Klein Tank et al., 2002]. The combination with a duplicate-removal procedure and with a quantile matching homogenization technique, analogous to the method presented by Squintu et al. [2019], allows the production of more reliable long temperature series, which are a fundamental tool for the estimation of indices and trends.

3.2 Data

The ECA&D collects data from thousands of stations spanning the European and Mediterranean domain. Temperature records play a key role, with about 3000 stored series per element (minimum, maximum and mean) (status January 2019, see table 3.1) which have been previously homogenized within the work presented in Squintu et al. [2019]. Figure 3.1 shows the spatial distribution of homogenized non-blended (OriHom, see subsection 3.2.1) series of average temperatures (TG). Here the size of the dot represents the length of the series and the color coding the starting year. This dataset clearly has a high potential for generating longer series, especially in the denser areas, and a blending procedure was developed to construct these augmented series. The procedure works by replacing missing data in the original non-homogenized series with measurements from surrounding stations and extends it to current times, when possible, with synoptic messages. A limitation of this approach is reflected in the production of duplicates of time series: in the case of a relocation, a new series is constructed joining donated data from the discontinued station to the receiving operational station *and vice versa*. Although the blended series are a fundamental tool for the development of the E-OBS gridded data-set [Haylock et al., 2008; Van den Besselaar et al., 2011; Cornes et al., 2018], duplicating time series in the input to this gridded data-set over-emphasizes the corresponding information. Thus an additional step to avoid any duplication is needed and has been included in this study.

3.2.1 Naming convention

The current ECA&D blended series (from now on referred to as *OldBlend*) were generated from the original non-homogenized series (*OriNonHom*) of ECA&D. The new blended series (*NewBlend*) are the result of the updated blending procedure described in the current study, which takes as input the homogenized original series (*OriHom*). These, as introduced by Squintu et al. [2019], are the result of two iterations of homogenization which remove large breaks. A third run of break detection revealed the presence of only minor inhomogeneities, which motivated not to further adjust in order to prevent overcorrection and preserve resemblance to the original data.

Finally, the homogenized version of the blended series will be referred to as *NewHomBlend*. The method described in this paper is applied to daily minimum temperature (TN), daily maximum temperature (TX) and daily average temperature (TG).

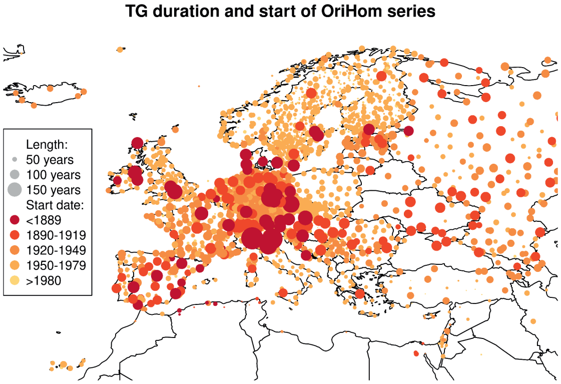


Figure 3.1: *Available TG homogenized series (OriHom, see section 23.2.1) in ECA&D. The size of the dots indicates the length of the series while color coding indicates the start year.*

3.3 Methods

3.3.1 The modified blending procedure

The construction of a NewBlend is performed for each station individually. Nearby stations around a target station are identified in a radius of 12.5 km and with a maximum of 25 meters of difference in altitude. These include the stations providing synoptic messages. This distance is the result of a pragmatic choice, based on the observation that mutual distances between stations affected by relocation tend to be below such threshold. All the OriHom data belonging to the target series and to the neighbouring series, which can be more than one for each station, are involved in the blending.

The series with the most recent data of the target station is called the *basis* series, while other OriHom from the same and neighboring stations are defined as *donating* series since they donate their data to fill possible gaps in the basis and to extend it to both earlier times and to more recent times (with GTS data).

In this process, illustrated in Figure 3.2, each day between the earliest and the latest date in the available series is considered individually. If the daily value is missing in the basis, the donating series are checked for availability for that date. The series which donates the datum is chosen according to a hierarchy based on distance from the basis, i.e. if the closest donor has a value on that date, it is selected, otherwise the second closest is checked and so on. In any case priority is given to validated series over synop values; the latter are used only when no alternative data are available and updates from the data providers are lacking. Integration with synop data is performed only if the last validated data lies in the 10 years preceding the current date [ECA&D Project Team, 2012]. In this step the introduced data are not adjusted for elevation or changes in the surroundings.

Figure 3.2 illustrates that the blended series is composed of segments that come from different series. The length of such segments might vary from several years to single days. Furthermore, contributions from a specific donating series can be fragmented (filling more than one gap in the basis series), making the homogenization more complicated.

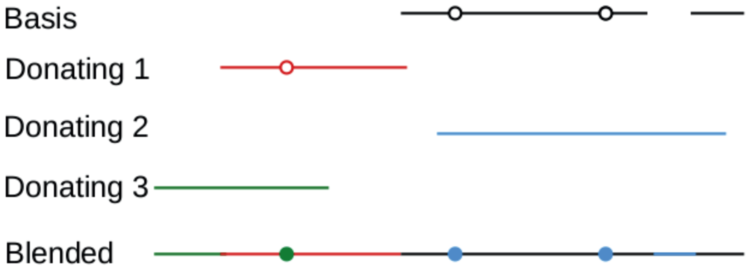


Figure 3.2: *Schematic diagram to show how data belonging to basis and donating series contribute to the creation of a long blended series. Lines indicate continuous segments, empty dots represent missing data. Missing data that have been infilled are indicated by full circles having the color of the corresponding donor. Donating series are sorted according to their rank.*

3.3.2 Avoiding duplicates

The described blending procedure determines the presence of identical segments of data, due to the mutual exchange of data that can occur between two neighboring stations. The presence of duplicates has the effect of giving double weight to certain records in the construction of gridded datasets like E-OBS.

The avoidance of duplicates in the new method is illustrated by the example of the area of Berlin (Figure 3.3). Initially all neighboring stations of Berlin-Tempelhof (ECA&D ID: 2759) are considered. For any of the neighbouring stations, it is checked if additional surrounding stations can be included. The process is iterated until no more stations are in the radius of 12.5 km of any of the already selected stations, determining a cluster of stations that have donated data to each other. The stations in the cluster are ranked giving priority first to those with OriHom which are still updated and then to those with the longest duration. According to these criteria Berlin-Tempelhof is ranked first. Among its donating series, the one which starts the earliest is the series of Berlin-Dahlem (ID: 41). This series is not updated anymore and its data are used to extend Tempelhof and Berlin-Tegel (ID: 4005) to earlier times. At the same time the blending procedure found data with ID 2579 to complete the series with ID 41 to current times. The portion of 41 before 1948 is used in the blending of station 2759, thus the values of this time interval are removed from 4005 and 41, since they have lower rank. The same process is applied for the segment between 1948 and 1963, which is present in 4005 and thus removed from 41. Stations Berlin-Buch (ID: 4529), Berlin-Marzahn (ID:4561) and Berlin-Schonefeld (ID:4570) are part of the cluster but their data cannot be used for the extension of 2579 nor can they profit from data from each other or from the lower ranked station since their distance to other stations is larger than the pre-set threshold. As a result of this process, no daily measurement appears in more than one blended series.

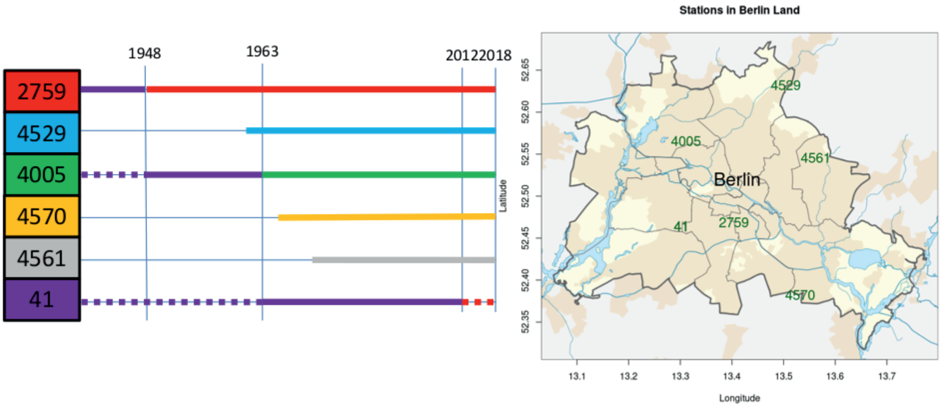


Figure 3.3: *Diagram and map representing the removal of duplicate records from blended series in the area of Berlin. Colors identify the series that contribute to the blended series. Dashed lines indicate data removed to avoid the presence of duplicates.*

3.3.3 Homogenization of the NewBlend series

In contrast to the OldBlend series, in which data from a neighboring station were simply inserted to fill gaps or to extend the record, the data of the NewBlend is adjusted to avoid the introduction of inhomogeneities. Although the generation of NewBlend series is applied to all elements available in the ECA&D database, the homogenization is applied to NewBlend temperature series only and is based on the probability density functions of the respective OriHom series. Methods to homogenize time series with a daily resolution for other elements are currently in an experimental phase and are outside the scope of this paper.

The calculation of adjustments is based on the quantile matching method documented by Squintu et al. [2019] which is based upon earlier work of Trewin [2013]. This consists of the comparison of the distributions of temperatures of the basis and of the donating OriHom. Data related to each month are considered separately over a period of at least 5 years from both basis and donating series. The records are divided into bins, each corresponding to the quantiles from 5 to 95. Sequences of values related to the quantiles are obtained, one for the basis and one for the donating series. These are subtracted from each other, in order to obtain an estimate of the adjustments. Since a climate change signal may be present in the difference of these quantiles, a comparison against parallel segments extracted from surrounding stations, which are used as references, is performed to remove this climate-related signal. More details are provided in the appendix of Squintu et al. [2019] and in Trewin [2013]. The reference series are selected among all the available OriHom in a coordinate box. This box always spans 6° in latitude, while the range of longitude varies according to the latitude in order to cover the same direction in the West-East direction that the box covers in the North-South direction. Each reference allows the calculation of an estimate of the adjustments. Finally, considering the set of estimates, the median is taken as the final correcting factor.

The adjustments are applied to each datum that appears in the blended series and which does not belong to the basis, according to its donating series, month and quantile.

3.4 Results

The procedure described above is applied to all the OriHom temperature series of ECA&D. Three case studies are reported in this section. The first (Geneva) illustrates the improvements on to the extremes indices, the second (Rheinstetten) represents an example of a series with discontinuous blending contributions from donating series and the third (Siauliai) explains how the GTS data have been treated and adjusted. Following these examples, an analysis of the effects on the whole data set is reported.

3.4.1 Case study: Geneva, Switzerland

Among the several examples of stations that have been relocated, one of particular interest is the series of TN from Geneva. The station in Parc de l'Observatoire was decommissioned in 1961. This station was likely influenced by nearby buildings and trees and by the Urban Heat Island Effect (see position in the map of figure 3.4). Simultaneously, the station in the open fields of the Cointrin Airport was established at a distance of 5.8 km and 15 meters higher than the old station. In the blending step these two OriHom are joined and an adjustment is required to avoid the introduction of an inhomogeneity.

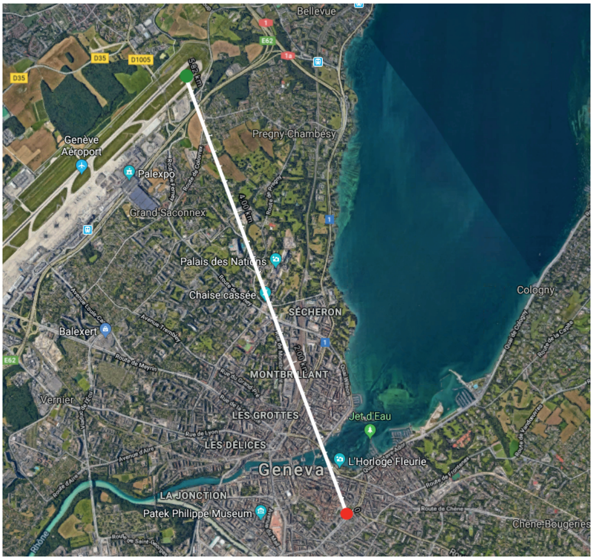


Figure 3.4: *Map of Geneva (courtesy of Google Maps) with the location of the station Parc de l'Observatoire (south) and Cointrin Airport (north) connected by the white line.*

As expected [Tuomenvirta, 2001; Böhm et al., 2010; Yang et al., 2013], the (indirect) removal of the urban heat island and, with lower impact, the elevation difference introduces a cooling factor to the NewBlend. A clear step-like pattern (black lines in figure 3.5) is observed when the series are joined without adjustment, while the red lines show the blending of the series using the homogenization adjustments. The average effect of the homogenization on annual means of TN (figure 3.5, top panel) is -1.3°C, while for the annual 10th percentile (bottom panel) the adjustments are stronger, -1.9°C. This shows that corrections based on mean values, in this case, underestimate adjustments for the more extreme values. This example illustrates how the combination of blending and homogenization generates a long and homogeneous series where the non-climatic effects of site location have not been adjusted.

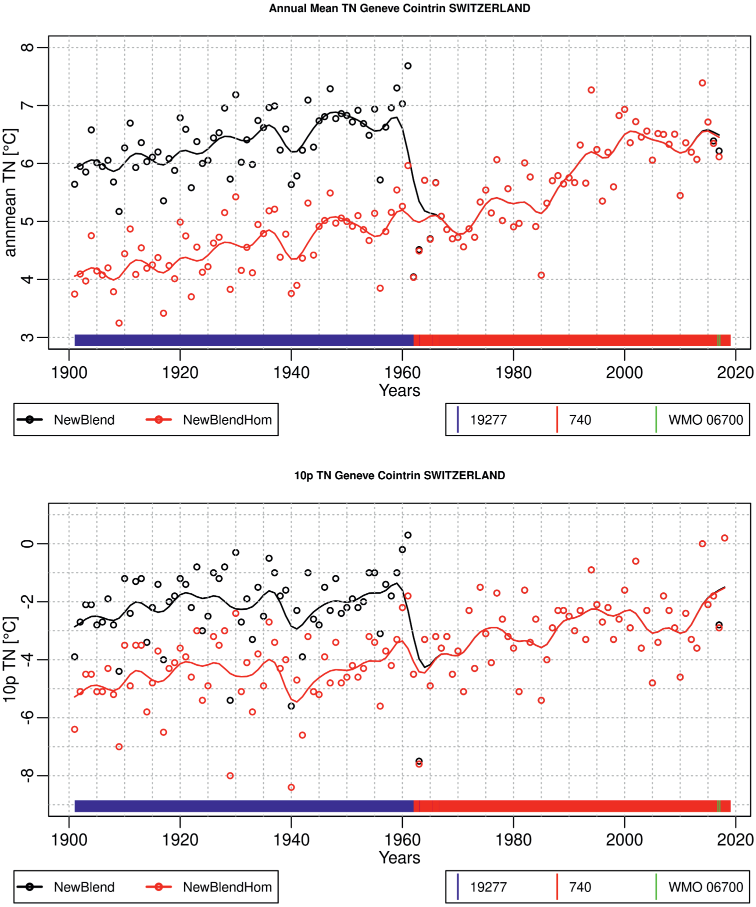


Figure 3.5: *Top panel: annual means of minimum temperatures of the blended series of Geneva. Dots stand for yearly values while the line is a gaussian weighted running mean. Black items represent the NewBlend, red items represent the NewHomBlend. Contributing segments are reported in the sequences of bars (one per day) on the bottom of the plots: the blue sequence indicates the period covered by Geneva Observatoire (ID:19277), red indicates Geneva Cointrin (ID: 740), and green indicates sparse contributions of the closest synop series (WMO 06700). Bottom panel: same as above but for the annual 10th percentile*

A final check has been performed to validate the results of this process against the homogenized temperatures provided by MeteoSwiss. These are monthly values of TG, thus the average of the 12 values has been compared with the average of the 365 daily values of the TG NewHomBlend series of Geneva. The strong agreement between the two series (see figure 3.6) is confirmed by the low RMSE: 0.15°C.

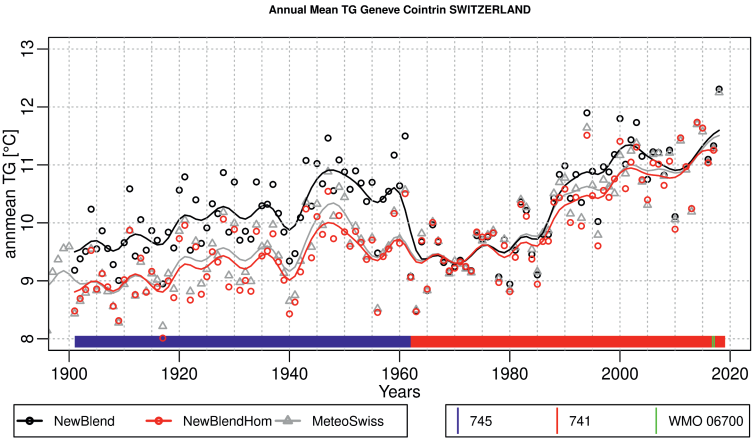


Figure 3.6: *Top panel: annual means of mean temperatures of the blended series of Geneva. Same color code as figure 3.5 with the addition of the annual average of the monthly homogenized values provided by MeteoSwiss (grey lines and dots).*

3.4.2 Case study: Rheinstetten, Germany

The meteorological station at Rheinstetten airport, Germany, was in operation between January 1948 and 1984 and again from 2009 onward. Data from this station are blended with data from the station Karlsruhe, which is located in a park approximately 8 km away from Rheinstetten (figure 3.7) and 4 meters of elevation lower. This urban station has an uninterrupted record running from 1876 to 2008, allowing the series of Rheinstetten to be extended further back in time and to fill the gap between 1985 and 2008. Figure 3.8 shows how in both periods the temperature in the Karlsruhe station appears to be warmer (as expected for a station in an urban environment) than the rural station. The applied adjustments convert these data to values that are more consistent with the basis.

The structure of this blended series is a representative, but rare, example with multiple relevant contributions from the same donating series. More frequently it is observed that multiple contributions from the same donating series consist of a main segment and several individual or short sequences of donated data that fill sparse missing values in the basis. During the homogenization step each segment is adjusted according to the

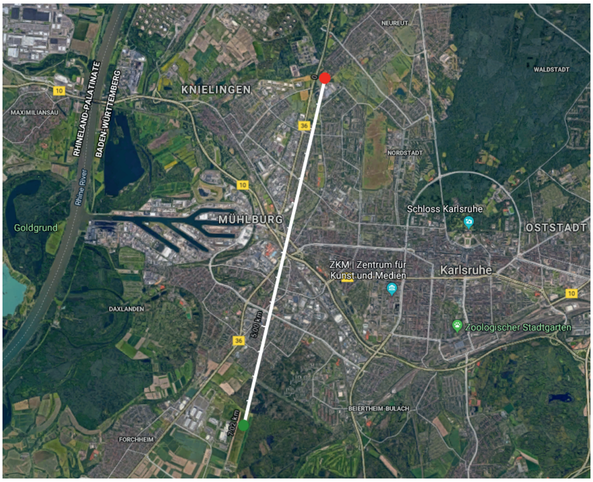


Figure 3.7: *Map of Karlsruhe and Rheinstetten (courtesy of Google Maps) with the location of the station Karlsruhe (north) and Rheinstetten airstrip (south) with the white line connecting the two stations.*

statistical features of the whole donating series. Thus the quantile adjustments are the same for all segments belonging to the same donor. Calculating separate adjustments for each segment would require arbitrary choices: determining which criteria to follow for splitting the donating series would make the process subjective. Furthermore, such a decision would imply the development of a more complicated algorithm which would be applied to a very small portion of the dataset (about 0.3%).

This motivated to apply as adjusting factors to blended data before 1948 and between 1984-2000 values which are drawn from the same set of quantile-based adjustments. Since the temperature values in the record before 1948 are generally lower than those in the latter period, the adjustments for the early period will sample the lower quantiles more than those from the more recent segment.

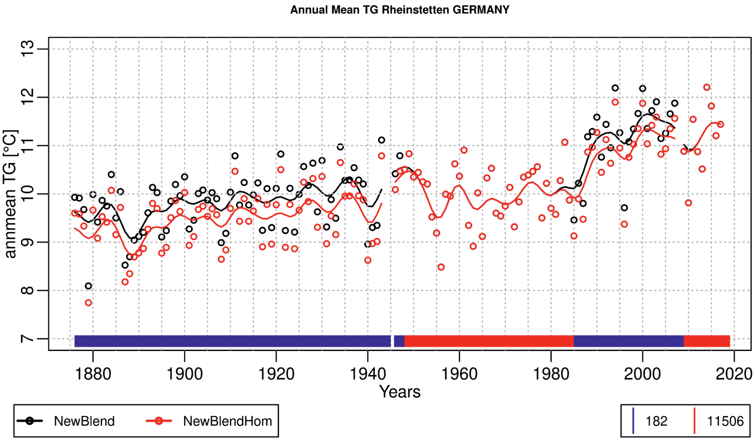


Figure 3.8: *Annual mean of the blended TG series of Rheinstetten. Same color code as figure 3.5. Red bar indicates data from Rheinstetten (11506), blue bar indicated data from Karlsruhe (182).*

3.4.3 Extending the series with homogenized synop data: Siauliai, Lithuania

Synop data are used for a maximum of 10 years to extend to present-day those series that are not regularly updated [ECA&D Project Team, 2012]. The daily TN and TX values are based on 12 hour periods (from 6 to 18 UTC for TX and from 18 to 6 UTC for TN, local time). This implies that synop values for TX (TN) are equal or underestimate (overestimate) the 24 hour-based TX (TN) values. Daily mean values (TG) are calculated as a simple average of TX and TN. Therefore this system is likely to introduce systematic biases on the last portion of the series and thus to introduce inhomogeneities [Van den Besselaar et al., 2012]. The quantile matching approach is used to adjust these biases, making synoptic data more consistent with data from the surrounding stations.

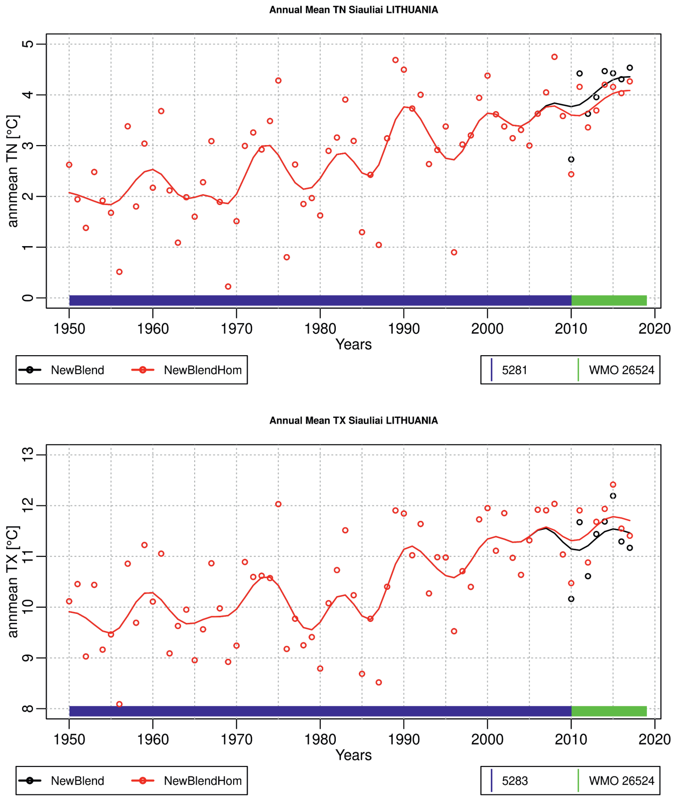


Figure 3.9: *Example of homogenization of GTS contributions to the series of Siauliai, Lithuania, for TN (top) and TX (bottom). Color code is the same as figure 3.5.*

It is important to stress that synops are appended and corrected, but not used as references; their statistical features are not transferred in any case to the surrounding series. An example of the correction of synops, for the TN and TX series of Siauliai, Lithuania, is shown in figure 3.9. As expected, the corrections applied to synops have low amplitude and different sign for maximum and minimum temperature. While this behaviour has been observed on almost all the most relevant contributions of GTS data, the histograms of adjustments applied to synops (figure 3.10) show that the negative adjustments dominate for minimum temperatures, while for maximum temperatures the distribution is centered around zero. This difference may be due to the fact that the probability of recording maximum temperature outside the 6-18 interval (local time) is lower than the probability of having minimum temperature occurring outside the 18-6 interval [Van den Besselaar et al., 2012]. Secondly the location of the station providing the synops series does not always coincide with the basis series, a fact that might interfere with the systematic biases introduced above.

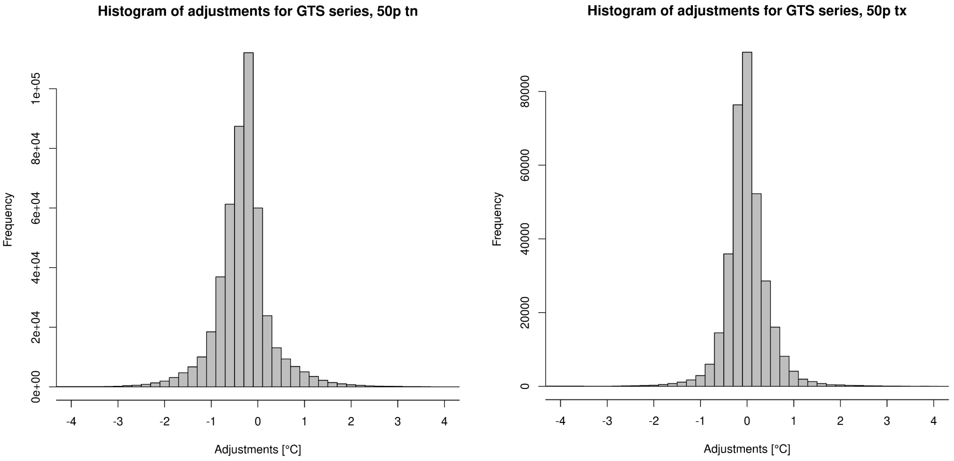


Figure 3.10: *Histograms of adjustments applied to the 50th percentiles of GTS series for TN (left) and TX(right).*

3.4.4 Application to the whole dataset: statistics

Table 3.1 describes the effects on the whole dataset. Considering the average lengths of OriHom (taking the longest for each station) and NewHomBlend, no considerable changes are observed (around +2 years). The transplant from a series to another together with the removal of duplicates explains the low increase. The most relevant change is observed in the distribution of the length of the series. This is illustrated in the histogram of figure 3.11 which highlights the shift towards larger values (thus longer series) confirmed by the increase (around +16%) of very long series (more than 60 years), see table 3.1. At the same time stations with a very low amount of data slightly increase their number. These series have donated most of their data to their neighbors.

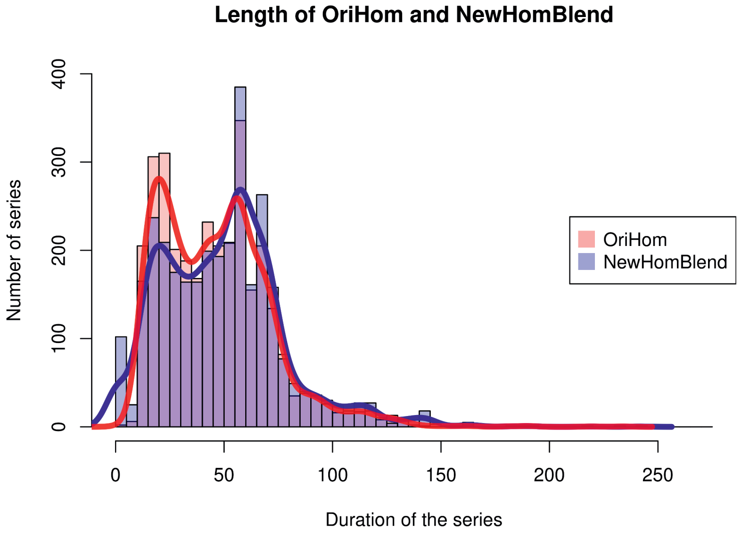


Figure 3.11: *Histogram and density function of the length of series before (OriHom) and after (NewHomBlend) the blending and the homogenization of the blended series. Red (Blue) items represent the original (final) dataset.*

Table 3.1: *Statistics regarding length of series. In the case of OriHom, only the longest series for each station is considered.*

	OriHom			NewHomBlend		
	TN	TX	TG	TN	TX	TG
Number of series	3199	3201	2710	3110	3114	2673
Number of series longer than 60 years	824	821	866	927	914	900
Average length	46.41	46.00	52.31	48.70	48.11	53.98

3.4.5 Effects of modified blending on trends

The results presented in sections 3.4.1, 3.4.2 and 3.4.3 indicate the power of the coupled blending and homogenization on individual series. Nonetheless, the correction of anomalous trends and the pursuit of a stronger spatial consistency of the trends is one of the primary tasks of the blending procedure.

Figure 3.12 compares the results for NewBlend and NewHomBlend. The trends displayed in those plots are for the annual 90th percentile of TX and are calculated using Sen's Slope method [Sen, 1968], which is more robust than linear regression; the significance of these trends is estimated using the Mann-Kendall test. Only the trends of those series with at least 80 years of data between 1911 and 2010 are considered, in order to focus on long term phenomena.

In the upper panel a lack of order is evident in the distribution of trends, with few apparent outliers. For example, Girona (Spain) 0.70 °C/decade, Milan (Italy) -0.20 °C/decade and Uccle (Belgium) -0.16 °C/decade. These is a clear sign of the presence of gross inhomogeneities. At the same time, less evident anomalies alter the spatial consistency, since the amplitude of such signals is comparable to those of the climatic trends. The disappearance of the extremely high and extremely low trends and the more consistent spatial patterns indicate the improved quality of the NewBlendHom dataset.

The map in the bottom panel of figure 3.12 shows the difference between the trends of NewHomBlend and NewBlend for each station. The difference of two trends is considered significant if the 95% confidence intervals of the two subtraction terms do not overlap. The absence of any pattern in such a plot indicates that the modified blending approach is neutral and does not favor an increase (decrease) of temperature trends over Europe, which would correspond to a predominance of orange/red (blue/purple) circles in the bottom panel of figure 3.12.

The median of the trends over Europe of NewBlend and NewHomBlend do not show relevant changes (0.14°C/decade to 0.15°C/decade for the 50th percentile of the three elements), i.e. when looking at the whole distribution, the differences in sign of the corrections compensate each other. Nevertheless changes in trends of the individual series are not negligible [Tuomenvirta, 2001]. The specific features of each area are filtered-out by the summarizing calculation (as mean, median, percentiles) due to the heterogeneity [Donat and Alexander, 2012; Li et al., 2016] of the dataset and to its non uniform density.

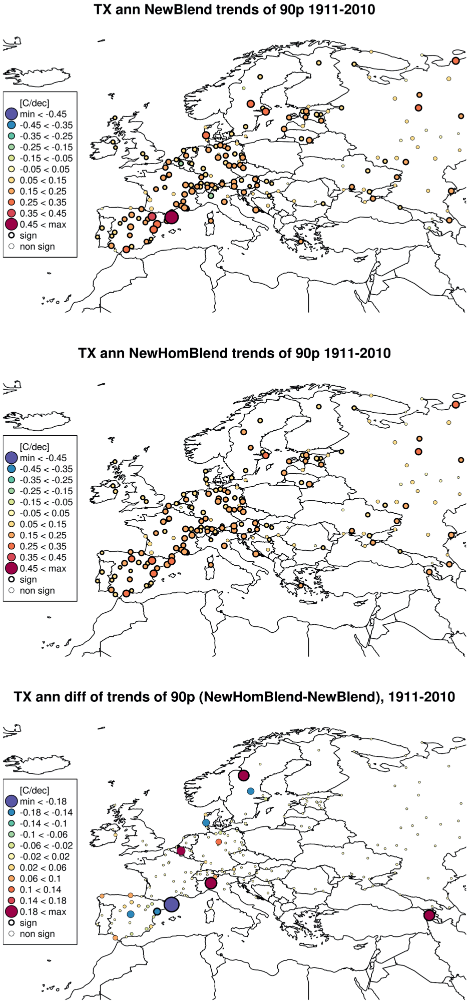


Figure 3.12: *Observed trends on 90th percentile of TX on the non-homogenized (NewBlend, top-left) and on the homogenized (NewHomBlend, top-right) blended series. Bottom panel represents the difference (hom-ori) of these trends for each station.*

3.4.6 Assessment of trends

The geographical patterns of temperature trends can now be assessed without being affected by the identified and introduced inhomogeneities. The quantile matching method provides different adjustments for the lower and the higher quantiles, and hence trends in the 10th and 90th percentiles of temperature are evaluated in this section. The maps in figure 3.13 show how the calculated trends vary over Europe. Winter (December, January and February) minimum temperatures (left column) show large trends in eastern Europe and in the Alpine region, especially for the lower percentiles. In the bottom panels changes in the shape of the distribution have been inspected by subtracting trends of 90th and 10th percentiles from each other. For winter minimum temperature, a narrowing of the probability distribution across the Eastern Mediterranean (Alps, Balkans, Ukraine) and Baltic Area is observed. This might be related to a decrease in the snow coverage of the areas during winter which disproportionally affects the *cold* tail of the distribution. At the same time Atlantic Regions and Arctic Russia show a slight increase in the distribution width. In contrast, summer (June, July and August) maximum temperatures (right column) have larger trends across southern Europe. While this increase has almost the same amplitude over the Iberian Peninsula, the results for Central Europe and Northern Italy are affected by steeper trends for the 90th percentile, which implies that the distribution of summer maximum temperature is becoming wider in these latter areas. These are the areas where the increase in intensity and duration of summer heat waves has had the highest impact, resulting in a larger trend on the *warm* tail of the distribution. These results are strongly related with the increase of frequency of extremely warm events, confirmed by several works on heat waves over the Mediterranean area [Della-Marta et al., 2007; Simolo et al., 2010; Yosef et al., 2018].

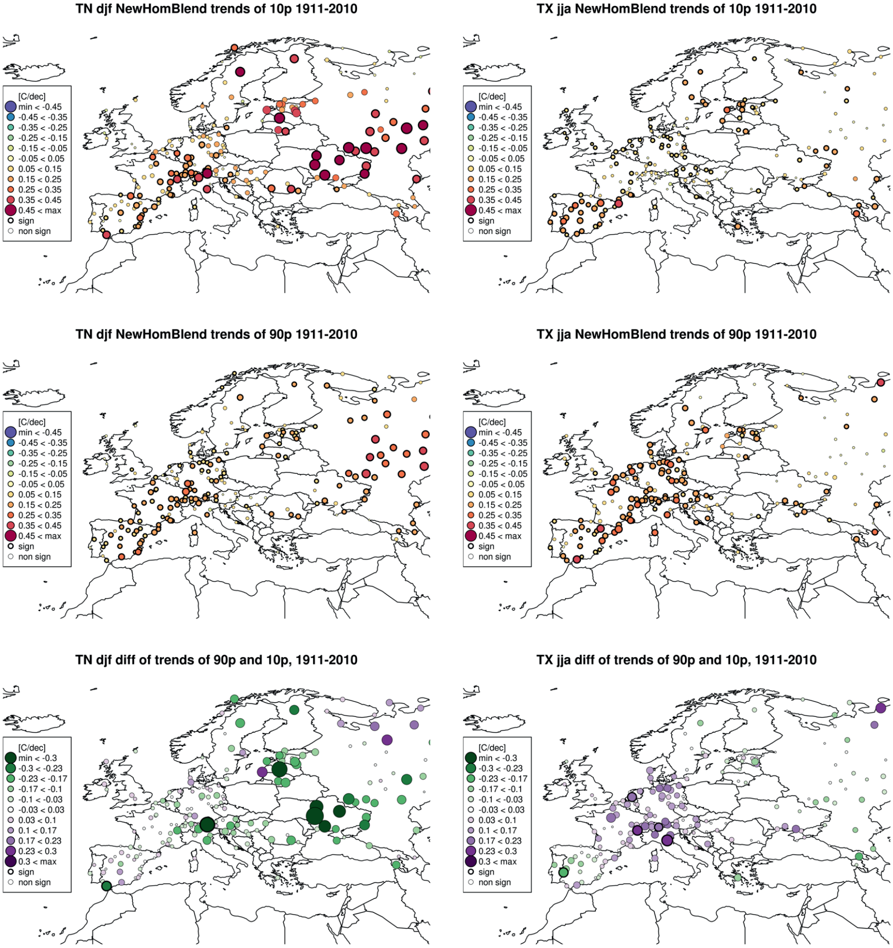


Figure 3.13: *Maps of trends of NewHomblend of winter TN (left) and summer TX (right) for 10th and 90th percentiles (top and center row) for the 1911-2010 period. Bottom row shows the difference between the trends for 90th and 10th percentiles: green (purple) represents narrowing (broadening) of the distribution.*

3.5 Summary and Discussion

In this study the homogenization procedure for temperature series documented in Squintu et al. [2019] is coupled with the blending procedure that is used operationally in the ECA&D [ECA&D Project Team, 2012]. The aim of the blending is to produce series that are as long and as complete as possible, facilitating climatological assessments. Long and homogeneous temperature records are constructed and trends in these records are compared against trends based on records that are constructed using the traditional method.

The most common situation that results in inhomgeneities is the relocation of a station from the city to the airport or to the countryside, which is expected to introduce a cooling signal in the resulting blended series. Furthermore, several other factors might also affect the series, including changes in the instruments and changes in the way the daily value is determined. Another specific issue that has been addressed is the extension of the series with synoptic messages. The inhomogeneity introduced when combining validated data with these synoptic data [Van den Besselaar et al., 2012] is also corrected.

The approach is illustrated using examples of three locations in Europe. A trend assessment comparing the new and the traditional method shows significant changes (variation in trends on 90th percentile of annual TX above 5% or below -5%) in 22% of the stations (figure 3.12).

With this new development, more reliable trend assessments on the extreme parts of the temperature distribution can be made as unrealistic outliers are removed. One example demonstrating this is the difference in trends in the 10th and 90th percentiles of seasonal temperature records. Dramatic differences between Eastern (warm trends in winter TN) and Western Europe (warm trends in summer TX) have been observed. In addition, the difference between trends in the 90th and 10th percentile of seasonal temperature has revealed a narrowing of winter TN distribution over South-East Europe and a widening of summer TX distribution over Central Europe. This highlights the importance of conducting regional analyses of climate change impacts.

Considering the results of the modified blending procedure for temperature, additional actions are needed in the future for its further development. One of the most ambitious ideas is the use of measurements of other meteorological variables (such as solar radiation, humidity or wind speed) to better interpret the differences among the stations that donate data to the blended series. The availability of a dataset with such long and homogeneous series will be a starting point for further studies on the characteristics of the warming climate of Europe.

The new blended and homogenized series will be used as the basis for a future version of E-OBS [Cornes et al., 2018] and should serve as a reliable benchmark for comparisons against computed trends in climate models.

All original data as provided by the ECA&D data providers, as well as the blended homogenized data, will be made available where permission to do so is given.

Acknowledgements

We acknowledge the data providers in the ECA&D project (www.ecad.eu), in particular Meteo Swiss for the cooperation in the case study of Geneva. Funding has been received from the EU FP7 Collaborative Project UERRA (Uncertainties in Ensembles of Regional ReAnanalysis), Grant agreement 607193 and the EU H2020 EUSTACE Project, Grant agreement 640171.

3.A Quantile Matching: brief description of the procedure

The quantile matching procedure is described in detail in Squintu et al. [2019], which has been developed from previous work by Trewin [2013]. As discussed in the main text the homogenization procedure applied to the NewBlend series takes as input the OriHom series that have contributed to the creation of the blended series itself. The latest ending series (excluding GTS) that belongs to the target station is called *basis*, $\mathbf{B}$, while all the other series (from the same and from neighbouring stations) will be referred to as *donating* series. The procedure makes use of a set of OriHom series located in the surrounding (within a coordinate box whose width varies according to latitude, see main text).

Adjustments are calculated considering individually each donating series, $\mathbf{D}_i$. For each series the first step is the definition of the reference list. At least three reference series, having at least five years of overlap and a correlation of 0.75 or higher with the basis and the considered donor, are required to proceed with the homogenization of the donated data; in the case of high availability the most correlated reference series are selected. For each reference $\mathbf{R}_j$, $\mathbf{R}_j^{\mathrm{B}}$ and $\mathbf{R}_j^{\mathrm{D}_i}$ are defined as the portions of the reference which are in overlap with, respectively, the basis and the donating series.

Adjustment calculation is performed for each month separately, including adjacent months in order to reduce the noise. The absolute temperature measurements related to the selected months of $\mathbf{B}$,$\mathbf{D}_i$,$\mathbf{R}_j^{\mathrm{B}}$ and $\mathbf{R}_j^{\mathrm{D}_i}$ are sorted in ascending order. The 5th , 10th , ..., 90th , 95th quantiles are selected, generating 4 quantile sequences ($\mathbf{b}_{q,m}$,$\mathbf{d}_{i,q,m}$,$\mathbf{r}^{\mathrm{B}}_{j,q,m}$,$\mathbf{r}^{\mathrm{D}_i}_{j,q,m}$).

The adjustments for $\mathbf{D}_i$, related to reference $\mathbf{R}_j$ for the month m and the quantile q are then calculated as:

$$\mathbf{a}_{i,j,q,m} = (\mathbf{b}_{q,m} - \mathbf{d}_{i,q,m}) - \left(\mathbf{r}^{\mathrm{B}}_{j,q,m} - \mathbf{r}^{\mathrm{D}_i}_{j,q,m}\right) \tag{3.1}$$

This process is iterated for each reference j=1,...,r.

At this point each value (v) which has been donated by $\mathbf{D}_i$, knowing its month, is adjusted looking, for each $\mathbf{R}_j$, at the quantile ($\tilde{q}_j$) it belongs to in the overlapping period between $\mathbf{D}_i$ and $\mathbf{R}_j$. Thus r estimates of the adjusted value are obtained:

$$\tilde{v}_j = v + \mathbf{a}_{j,\tilde{q}_j,m} \tag{3.2}$$

The final adjusted value is then calculated taking the median of the estimations:

$$\overline{v} = \mathrm{median}_j(\tilde{v}_j) \tag{3.3}$$

This process is iterated for each donating series that contributes to the blended series.

Chapter 4

Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset

Abstract

In this chapter four homogenization methods for daily temperatures (together with two variants) have been tested and compared. This has been performed constructing a benchmark dataset, where segments of homogeneous series are replaced with simultaneous measurements from neighbouring homogeneous series. This generates inhomogeneous series (tests) whose homogeneous version (benchmarks) is known. Two benchmark datasets are created: one based on series from the Czech Republic (with high quality, high station density and a large number of reference series) and one more challenging (missing segments, low station density and scarcity of references) with stations from all Europe. The comparison has been performed with pre-defined metrics which check the statistical distance between the homogenized versions and the benchmark. Almost all methods perform well on the near-ideal benchmark (maximum relative RMSE: 1.01), while on the European dataset, the homogenization methods diverge and the rRMSE increases up to 1.87. Analyses of the percentages of non-adjusted inhomogeneous data (up to 39%) and substantial differences in the homogenized trends helped identifying diverging procedural characteristics of the methods. These results add new elements to the debate about homogenization methods for daily values and motivate the use of realistic and challenging datasets in evaluating their robustness and flexibility.

This chapter was originally published as:

Squintu, A. A., van der Schrier, G., Štěpánek, P., Zahradníček, P., and Klein Tank, A. (2020a). Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theoretical and Applied Climatology*, pages 1–17

https://doi.org/10.1007/s00704-019-03018-0

4.1 Introduction

Homogenization of climatic data is a fundamental step in any climatological analyses. Human intervention on measuring stations induces sharp or gradual changes to temperature time series, affecting the reliability of the climatological analyses.

In the last decades several homogenization methods have been developed. While initially the focus has been on the adjustment of annual, seasonal or monthly series (thus mainly studying the changes on the first moments of the temperature distribution) in the last two decades several studies have focused attention on daily values and on the effects on extreme events, which have more effect on higher moments.

Homogenization procedures are divided into two phases: the break detection, that is generally applied to series that are aggregated to the yearly, seasonal or monthly time scale, and the adjustment calculations. The first step is not part of the focus of this paper. The latter is suitable to be adapted to a daily resolution and has been subject of a vivid debate. Several studies have stated that the use of metadata and the eventual subjective assessments and decisions are preferable for a thorough homogenization [Venema et al., 2013; Pérez-Zanón et al., 2015; Gubler et al., 2017; Delvaux et al., 2018]. Nevertheless the size of certain datasets and the need for objectivity and reproducibility require the use of completely automated processes [Venema et al., 2013]. The currently existing automated approaches for the adjustment of daily series range from methods that are based on the detection of a yearly cycle [Vincent et al., 2002], to the parametrization of the distribution [Caussinus and Mestre, 2004; Della-Marta and Wanner, 2006; Mestre et al., 2011], or to more heuristic approaches [Menne and Williams Jr, 2009; Štěpánek et al., 2009; Trewin, 2013; Squintu et al., 2019]. The different mathematical and statistical theories behind the methods and their various procedural aspects (e.g. separation between break detection and adjustment calculation, number of iteration of the process, different number of references, etc.) imply that the homogenized version of a inhomogeneous series will be different depending on the method that is used to homogenize it.

The comparison of the efficiency and capability of these methods is then a powerful tool for the detection of their strong and weak spots, contributing to the development of more capable homogenization systems.

In several cases the comparison among methods has been performed as a validation for newly introduced homogenization methods [Caussinus and Mestre, 2004; Menne and Williams Jr, 2005; Della-Marta and Wanner, 2006; Mestre et al., 2011; Trewin, 2013; Vincent et al., 2018]. On the other hand some outstanding studies have inspected the efficiency of automated, semi-automated and manual homogenization procedures [Venema et al., 2013; Domonkos, 2013; Li et al., 2016], focusing on break detection skills and on the performance on monthly data [Venema et al., 2013].

Every comparison of homogenization methods must ensure transparency and objectivity. The establishment of a benchmark dataset and the definition of clear metrics and evaluation parameters, determined before the examination of the results, give a fundamental contribute to this purpose.

The benchmark series are series that are meant to carry only the climatic signal, which can be either real or simulated, and are used as truth to be reproduced by a homogenization method. These records, which must be homogeneous and lack any quality issue, are intentionally perturbed in order to create fictitious inhomogeneous test series that will be the input of the evaluated homogenization procedures. The creation of the test series has usually been implemented with the insertion in the benchmarks of missing values, outliers, trends, noise and inhomogeneities [Menne and Williams Jr, 2005; Mestre et al., 2011; Williams et al., 2012; Venema et al., 2013; Domonkos, 2013; Lindau and Venema, 2016]. The temporal frequency and amplitude used for the insertion of such events is chosen following ad-hoc studies, in order to reproduce as much as possible realistic situations. An evolution of this approach considers the creation of ensembles [Domonkos, 2011; Vincent et al., 2018; Domonkos and Coll, 2017] of perturbed test series. This is done by varying amplitude and frequency of the breaks, allowing to test the efficiency of homogenization methods under particular conditions. Though such an approach is not able to reproduce all the possible artificial signals that might be introduced to inhomogeneous series [Vincent et al., 2018], giving more relevance to those signals that are easier to be studied and recognized. Moreover the events and phenomena that can affect a recording station are so various, complex and unexpected, that recent works have considered the opportunity to rely only on validated data (instead of simulated ones) for the creation of the test series [Venema et al., 2013; Gubler et al., 2017; Vincent et al., 2018]. This approach has been recently developed by Trewin [2013] who elaborated the idea of joining records from close-by stations. This method allows to test homogenization methods on real data (thus carrying all the possible signals). Nevertheless with this choice there is no knowledge on the truth (i.e. knowing the series as it was affected by only the climatic signal), since the benchmark, even after a preliminary homogenization and quality check, will never be completely free from artificial signals.

Whatever method is chosen for the creation of a benchmark dataset, the considered homogenization methods are applied to the test series with the aim to reproduce the benchmarks. The *distance* between the homogenized versions and the benchmarks, evaluated with transparent and clear metrics that must be defined in advance, indicates the performance of the homogenization. The most common used metrics involve calculations of relative RMSE (rRMSE) [Mestre et al., 2011; Venema et al., 2013; Domonkos, 2013; Trewin, 2013; Domonkos and Coll, 2017; Gubler et al., 2017; Vincent et al., 2018], changes in the trends of indices [Domonkos, 2013; Venema et al., 2013; Pérez-Zanón et al., 2015; Domonkos and Coll, 2017] and countings of number of days within fixed absolute thresholds around benchmark values (e.g. PD05, see section 4.2.4) [Trewin, 2013; Vincent et al., 2018]. Finally the calculation of network mean biases of the trends [Domonkos, 2013; Domonkos and Coll, 2017] are a good indication of systematic behaviours related to certain homogenization methods and provide a good general estimation of the performance. In the specific case of evaluation of homogenization methods of daily data it is important to focus on the effect on trends of indices that are related to extreme values [Trewin, 2013], such as 10p (10th percentile of a daily series calculated on annual basis) or 90p (same but for the 90th percentile).

Benchmarking and evaluating methods helps a scientific community to develop and get mature [Venema et al., 2013]. On one hand users which are performing climatological studies are assisted in the evaluation of how the methods fit their needs. On the other hand the researchers that develop homogenization software can improve the methods including new features and facing statistical and procedural challenges with more awareness. Furthermore this stimulates the debate on the comparison methods themselves, pushing in the direction of the development of more reliable system for the creation of benchmark datasets and more sophisticated metrics.

The aim of the present work is to evaluate the performances of the homogenization methods that have been considered and to highlight criticisms in their statistical and procedural features. An important focus will be given to the construction of a reliable benchmark and the description of clear metrics which are mandatory steps to assess solidity and transparency of any study of this kind. This analysis is limited to methods that can be applied to large datasets, thus completely automated. Several methods have been developed in the last decades ([Della-Marta and Wanner, 2006; Menne and Williams Jr, 2009; Štěpánek et al., 2009; Wang et al., 2010; Mestre et al., 2011; Trewin, 2013; Domonkos and Coll, 2017; Guijarro, 2018; Squintu et al., 2019]). In the framework of Copernicus Project (C3S.311a-Lot4) only the four listed below have been tested. The involvement of more methods is planned and will be implemented within future works.

4.1.1 Compared methods

The Quantile Matching, developed within the European Dataset Climate and Assessment (ECA&D), calculates adjustments via the comparison of the distribution of temperatures before and after a detected break. This happens taking the difference of similar quantiles (5th , 10th , etc.) in the two periods, identifying the changes in the distribution in the two target segments. The same comparison is performed on contemporary segments in a set of reference series, identifying which part of the difference is due to the climatic background. The obtained adjustment estimates depend on the quantile, on the month and on the corresponding reference series, their median value is then applied to the daily values. Homogeneous references are chosen among the highest-correlated surrounding homogeneous series, with threshold on distance and correlation. A series is corrected if at least 3 homogeneous reference series with at least 5 years of overlap are available. See [Squintu et al., 2019] for more details.

The DAP (Distribution Adjusted by Percentile) method [Štěpánek et al., 2009, 2013] works as well on empirical distributions. The process calculates the day-by-day difference between candidate and reference series (obtained as average of 5 highly correlated neighbouring series), such values are binned according to the percentiles of the candidate series. This is performed separately for the portions of the series before and after the break and iterated for each month individually, including adjacent months. Finally the differences of the results before and after the break are calculated and smoothed by a low-pass filter. For each datum before the break its quantile is determined via interpolation of the percentiles and the corresponding adjusting factor is then applied. The tails of the distribution receive a special treatment: if the correlation after correction does not improve by at least 0.05% (considering each month individually) the adjustment is not applied and the original data are preserved. Note that even if both QM and DAP are based on quantiles, the first one manages the reference series with a "pairwise comparison" approach [Menne and Williams Jr, 2009; Trewin, 2013] while the latter in this work calculates the average of the selected references, coherently with all the previous applications of DAP performed by GCRI. Finally there are several different procedural differences, as the low-pass filter and the correlation improvement threshold used by DAP.

The third tested method was developed by [Della-Marta and Wanner, 2006] and is widely known as High Order Moments (HOM). This works comparing the candidate with an overlapping reference series. A model is developed comparing candidate and reference after the break. The same model, applied to the reference segment before the break, is used to create a predicted series. The values of difference between observed and predicted are binned in deciles according to the quantile of the observed series. The values related to each decile are fitted with a local estimated scatterplot smoothing (LOESS) function, that determines a smooth set of adjustments depending on the quantile of the candidate series.

Finally SPLIDHOM [Mestre et al., 2011] considers as well only one highly-correlated reference series, which can also be inhomogeneous. Regressions based on cubic splines are calculated in order to describe the relationship between the candidate and the reference in the two overlapping periods before and after the break. The difference of these two models allows to determine the adjustments basing on the values of the reference before the break. The values of this last segment can be obtained from the candidate values with the inversion of the regression introduced above. Thus, starting from the value of the candidate before the break, the adjustment can be calculated knowing the regression relationships between the candidate itself and a reference series.

This last method, differently from the previous ones, doesn't make use of quantile bins, but it completely relies on regression methods. On the other hand, while HOM couples regression and quantile binning, Quantile Matching and DAP are solely based on empirical features.

HOM and SPLIDHOM have been tested twice, the first time with their original configuration and the second time using procedural approaches similar to DAP. For example the application of a low-pass filter with coefficients retrieved from a Gaussian distribution (more reliable than a moving average [Mitchell et al., 1966]) in order to smooth the adjustments and a threshold on the improvement of correlation have been applied. leaving unchanged those data that don't fulfill such constraint. Thus these variants keep the statistical aspects of HOM and SPLIDHOM, coupling them with the more conservative approaches typical of DAP.

4.2 Data & Method

4.2.1 Construction of a benchmark dataset

A benchmark dataset is a fundamental tool in the comparison of homogenization methods. It is based on so-called benchmark series (*benchmarks*), which must be homogeneous and lack serious quality issues. In this work the data of the benchmarks are partially replaced by homogeneous segments of data coming from surrounding series (*perturbers*), following Trewin [2013]. Such intervention generates test series (*tests*), which are inhomogeneous, but composed of homogeneous segments, and need to be adjusted. The homogenization of such test series, adjusting the perturbing segments, is performed by the selected methods with the help of a set of homogeneous reference series (*references*), whose density and data availability can change. The different homogenized versions are compared through established metrics.

The challenges of a benchmark dataset are in the inhomogeneities themselves, which can have various amplitudes, structure and timing. At the same time the data availability and the station density has a great influence on the performance of homogenization [Caussinus and Mestre, 2004; Domonkos, 2013; Gubler et al., 2017]. In this study two different benchmark dataset have been created for these reasons. The first one consists of a very dense dataset in the Czech Republic where 20 benchmark series are accompanied by more than 200 references. Here the methods can easily work and their results can be compared in favourable conditions. The second one aims at testing the capability of the homogenization methods to work in difficult conditions where the number of homogeneous reference series is low. This implies that test series are required to be used as reference for each others homogenization and that in some cases the homogenization is impossible due to data scarcity. In the following sections the two benchmark datasets are explained in detail.

4.2.2 The Czech benchmark dataset

The Czech dataset aims at testing the methods on a nearly-ideal situation where station density and series continuity is high and where the group of series to be adjusted represent a very small percentage of the whole dataset (less than 10%), while the remainder includes homogeneous series.

In such favourable conditions the statistical features of the homogenization methods can be compared independently from any thresholds on data availability and data quality that are used in the software scripting and reflect the views of the developer of the minimal conditions for which homogenization is meaningful.

The benchmark series (green dots in figure 4.1) span the 1970-1999 period and have been perturbed in the first 15 years with data coming from homogeneous neighbouring stations (black dots), so that the breaks are all set to the same date (1st of January 1985). More than 200 homogeneous reference series (purple dots) are provided to help the homogenization. The used series are all provided by CHMI and have been preliminarily checked in their quality and in their homogeneity with Proclim DB software [Štěpánek et al., 2009, 2013].

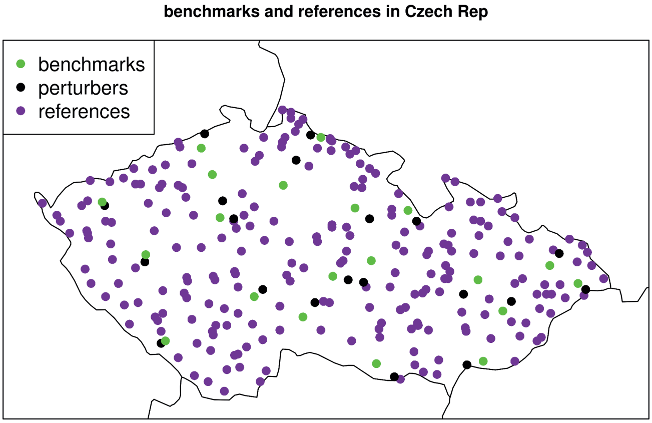


Figure 4.1: *Geographical distribution of benchmark and test series (green dots), perturbing series (black) and reference homogeneous series (purple dots). Benchmark series have same location for minimum and maximum temperatures.*

4.2.3 The European benchmark datasets

The European benchmark dataset is composed of original series extracted from the European Climate and Assessment Dataset (ECA&D, [Klein Tank et al., 2002]) and are based on actual observations that did not undergo homogeneity adjustments. Benchmarks are chosen among the set of already homogeneous series of the ECA&D. Their homogeneity is tested through the agreement of three tests (Prodige [Caussinus and Mestre, 2004], RHtest [Wang et al., 2007] and GAHMDI [Toreti et al., 2012]) according to the method documented by Kuglitsch et al. [2012] and Squintu et al. [2019]. All the series which have been assessed as homogeneous in the 1970-1999 period are considered to provide the benchmarks. For each of the potential benchmarks the neighbourhood (radius of 20 km and maximum elevation difference of 50 meters) is inspected in order to find series whose homogeneous segments will be used to perturb the benchmarks. A nearby series with at least 6 years of homogeneous data between 1970 and 1993 is found become eligible as a perturbing series (*perturber*). All homogenization procedures adjust the earlier part of the series consistent to their most recent part [Aguilar et al., 2003]. Hence the last six years of the considered period (1994-1999) are preserved for the data coming from the benchmark and will be used as basis for the calculation of the adjustments. The potential perturbing segments are checked to be sure they are not total or partial duplicates of the benchmarks, analyzing correlation and absolute mean of daily differences. If no neighbours following such constraints are found, the potential benchmark is discarded. This process is performed separately for minimum (TN) and maximum (TX) temperatures. In order to generate a dataset of moderate size, only 50 TN and TX benchmark series (among circa 70 series per each element) are chosen. This selection has given priority to those stations that provide series to both datasets (more or less 35). The remaining ones are chosen in order to approximately reproduce the same geographical distribution for both variables, with few interesting exceptions (Belarus for TN, Serbia for TX, see figure 4.2).

Once benchmark and its perturbers are identified, the test series are created, replacing the data of the benchmark with the data of the perturber. In case only one perturber is related to a certain benchmark, the whole (homogeneous) perturbing segment is used to replace the benchmark data. Thus the change point will be at the end of the perturbing segment. If more perturbers for the same benchmark are found, these have been combined in order to donate segments of at least 6 years of data. In case the identified perturbers are homogeneous all through the whole period, the change point has been set on the 1st January of 1985. This has allowed to create a sub-set of test series with simultaneous breaks, which is a great challenge for the homogenizers in current times [Venema et al., 2013]. Figure 4.3 illustrates the building of the test dataset. The generated change points (*breaks*), whose distribution can be observed in figure 4.4, have been transcribed, together with coordinates and elevations of the benchmarks and their corresponding perturbers.

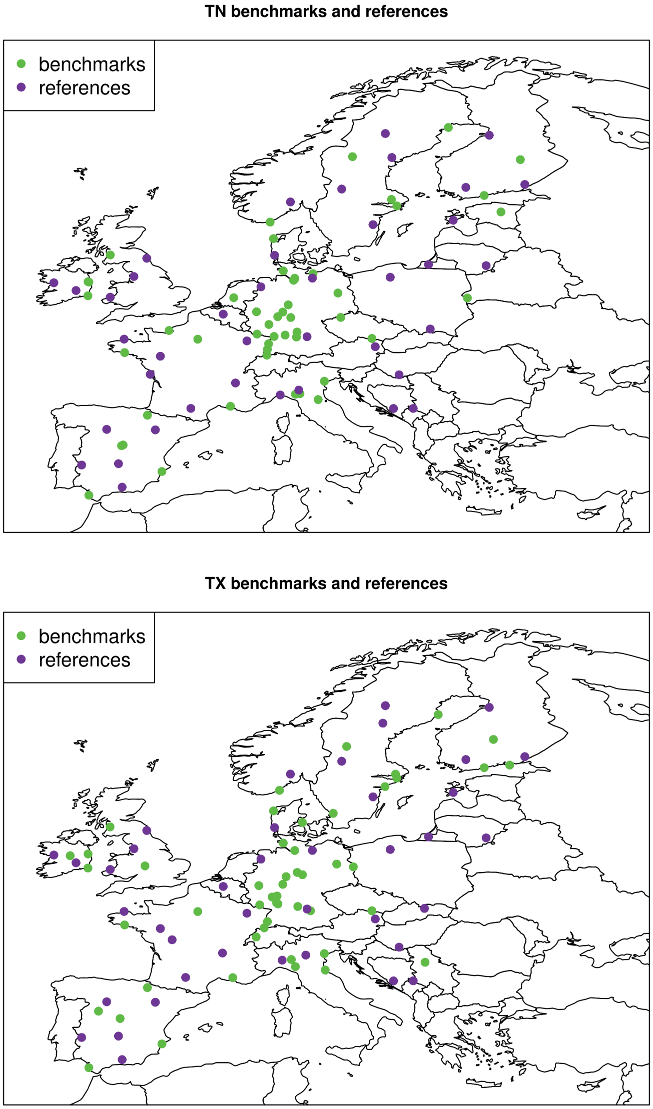


Figure 4.2: *Geographical distribution of benchmark and test series (green dots) and reference homogeneous series (purple dots) for minimum (top) and maximum (bottom) temperatures.*

The set of 50 test series for each variable has been integrated with 40 homogeneous series, chosen among the discarded potential benchmarks. The selection has been performed on a geographical uniform pattern. Since the percentage of homogeneous series observed in the ECA&D dataset is 35%, the goal has initially been to reproduce such proportion

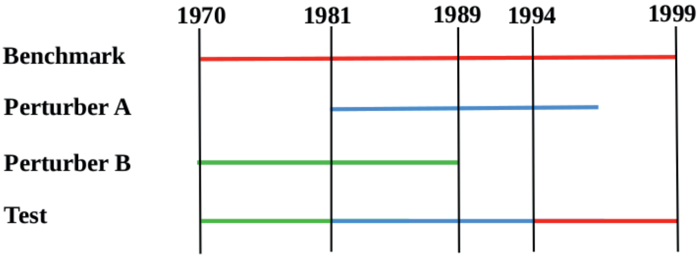


Figure 4.3: *Schematic diagram describing how a test series is generated. Each of the perturbing segments is required to be at least 6 years long and after 1994 only benchmark data is used.*

between homogeneous and non-homogeneous series. This choice would have caused a lack of series in some areas (such as Spain, UK, France) and would have lead to the possibility of not being able to homogenize them at all. For this reason the references have been selected so that each benchmark has at least 3 homogeneous references in its surrounding area (see figure 4.2).

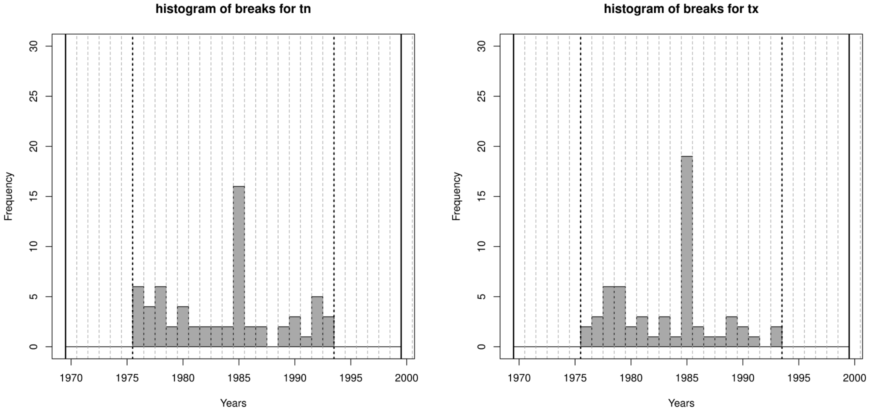


Figure 4.4: *Distribution of inserted breaks in the European Benchmark dataset, which derives from data availability and automatic combinations. The peak in 1985 is the result of the adding of simultaneous breaks (see text). No breaks are allowed before January 1976 and after December 1993 (dashed vertical lines), since each segment in the test must be at least 6 years long.*

Figure 4.5 shows the distribution of daily differences for each pair of test and corresponding benchmark for the Czech and European tests, and for both daily maximum and mimimum temperature. The amplitude of the breaks shows a strong variation, as demonstrated by the interquartile range and the 5th and 95th quantiles. Although the centres of the distribution tend to be more in the negative half of the plot, for almost all the series the distributions of the breaks cover negative and positive values. This indicates the need of particular care in the homogenization process and confirms the complexity of inhomogeneities.

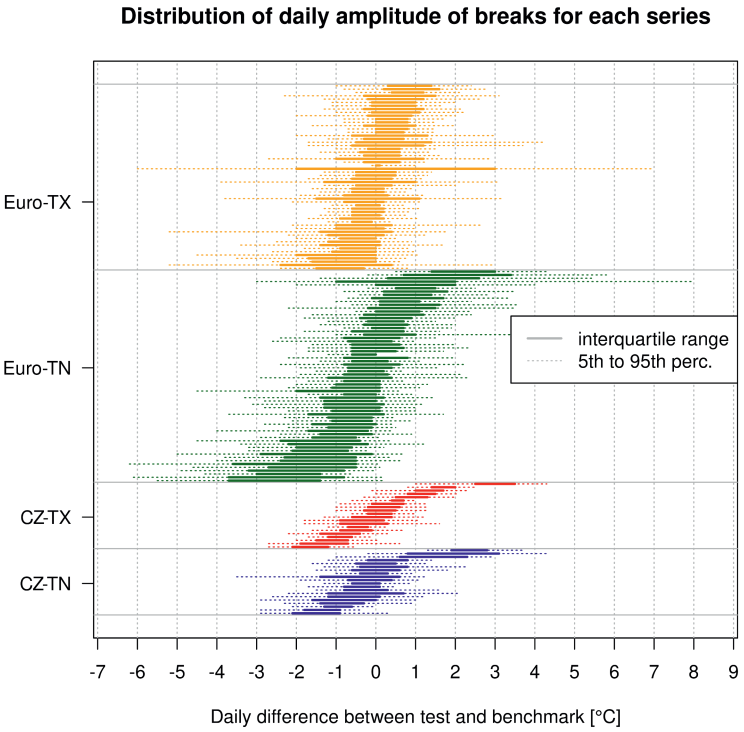


Figure 4.5: *Distributions, series per series, of the daily differences between test series and corresponding benchmark series. Thick line indicates interquartile range, while dotted lines stand for the tails of the distribution up (down) to 95th (5th) quantile.*

4.2.4 Definition of metrics

A reliable comparison of homogenization methods requires a set of metrics, defined in advance. The most commonly used metric is the relative RMSE (rRMSE) ([Mestre et al., 2011; Venema et al., 2013; Domonkos, 2013; Trewin, 2013; Vincent et al., 2018; Domonkos and Coll, 2017]) which allows to measure the distance between the homogenized and the benchmark version of a series. Then an average of the results on all the benchmark series and an analysis of their distribution allow to compare all the tested methods. A powerful metric, developed specifically for daily series, is the PD05: number of days where the day-by-day difference between the homogenized version and the benchmark is lower than 0.5 °C[Trewin, 2013; Vincent et al., 2018]. This is useful in defining a day by day accuracy of the homogenization, even though it may be positively biased in cases of breaks with low magnitude.

Even though these metrics give a general overview of the power of the tested methods, they lack the information about variability of the series and about the effect of the homogenization methods on extreme values and interannual variance, which are of great importance [Trewin, 2013]. The trends assessment, inspecting how the homogenized versions get close to the benchmark trend, are one of the metrics that allow a thorough comparison of the methods [Domonkos, 2013; Venema et al., 2013; Pérez-Zanón et al., 2015; Domonkos and Coll, 2017]. In this paper trends on 10^{th} , annual mean and 90^{th} percentile are compared.

The distance of the obtained trends from the one of the benchmak is measured with a homogenization indicator (*hom_ind*) defined by the following expression:

$$\text{hom\_ind} = 1 - \frac{hom - ben}{tst - ben} \tag{4.1}$$

where *hom* is the homogenized value (which can also be daily value, annual mean, annual percentile, etc.), *ben* is the value related to the benchmark (the goal of the process) and *tst* is the value related to the test series (the starting point).

Such indicator takes the following values:

- 0 when the adjustments, if any, have no effect on the series;
- 1 when the adjustment perfectly reproduce the benchmark value;
- above 1 when the series is over-adjusted;
- between 0 and 1 when the series is partially adjusted;
- negative values when the series is adjusted in the wrong direction;

This metric allows an estimation of the accuracy of the homogenization and is more powerful than a simple difference between homogenized and benchmark values that would be difficult to analyze in the general context, due to the specific features of each benchmark series. Nevertheless when the difference between *tst* and *ben* is close to zero, singularities appear, limiting the significance of the indicator in such eventualities.

In the case of the European benchmark dataset the scarcity of data may cause that one or more of the tested methods does not apply adjustments to the test series (or part of it). For this reason a preliminary evaluation of the number of non effective adjustments is made. The fraction of non-adjusted data is reported as ratio between the number of daily data where the difference between test value and the homogenized value is zero and the number of daily data that has to be adjusted (i.e. parts of the test series that belong to the perturbers). In this work the homogenization of a test series is considered fruitless when the percentage of non-adjusted data is above 80%. Such threshold is high enough to avoid considering fruitless series with a high amount of data whose calculated adjustment is actually zero or series with high percentage of missing data.

4.3 Results

4.3.1 Czech dataset

The introduced metrics have been used to evaluate the performance of the homogenization methods on the Czech dataset. As expected, due to the good quality of the dataset, all the methods managed to homogenize all the series in all of their parts. In the following section the colour code displayed in table 4.1 is used.

Table 4.1: Acronym, references and colours associated to the homogenization methods. Colours are used for figures from 4.6 onwards.

method	reference	acronym	colour
Test series (inhomogeneous)	see section 4.2.1	inh	black
Quantile Matching	[Squintu et al., 2019]	QM	red
DAP	[Štěpánek et al., 2009]	DAP	blue
HOM	[Della-Marta and Wanner, 2006]	HOMHOM	yellow
SPLIDHOM	[Mestre et al., 2011]	SPLIDHOM	green
HOM with 'DAP' parameters	see section 4.1.1	HOMDAP	purple
SPLIDHOM with 'DAP' parameters	see section 4.1.1	SPLIDHOMDAP	orange

Table 4.2: *Results of methods comparison on the Czech dataset. Underlined values are discussed in the text.*

dataset(CZ)	mean of RMSE [°C]		mean of PD05[%]		mean of hom_ind of trends of					
					annmean		10p		90p	
	TN	TX	TN	TX	TN	TX	TN	TX	TN	TX
Test Series	1.31	1.17	39.01	37.44						
Quantile Matching	0.96	0.73	52.79	65.48	1.37	1.08	0.71	0.81	1.09	0.63
DAP	0.96	0.74	51.74	64.89	1.03	1.12	0.83	0.70	0.90	0.62
HOM	0.99	1.03	50.59	62.82	0.86	1.01	0.77	0.68	0.73	0.42
SPLIDHOM	0.97	0.74	51.50	64.76	0.89	1.12	0.86	0.70	0.80	0.63
HOM (DAP)	0.98	0.74	50.86	64.38	0.99	1.10	0.59	0.51	0.79	0.47
SPLIDHOM (DAP)	0.96	0.74	51.82	64.93	1.09	1.12	0.88	0.71	0.81	0.48

First inspection is performed with the rRMSE, taking the benchmark version as reference, here lowest values indicate the best performances.

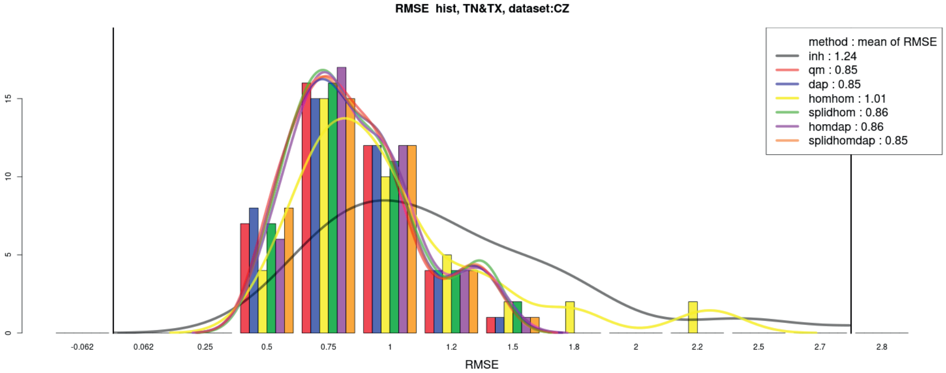


Figure 4.6: *Histograms and density plots of rRMSE for the comparison of methods on the Czech dataset.*

In figure 4.6 all the methods show similar behaviours except HOM (yellow line). The deviation of this method in comparison to other methods is confirmed in table 4.2, where HOM shows a high average of rRMSE for the homogenization of TX. The high value of mean rRMSE is due to the performance on two series, one of these can be observed in figure 4.7. Here it is clear how the results of HOM do not correctly reproduce the annual mean of the benchmark (grey thick line). The anomalous behaviour of HOM is confirmed by plotting the homogenization indicator (equation 4.1) applied to the yearly values (bottom panel) where all the methods (except QM) show to be over-adjusting the series (hom_ind>1), while HOM under-adjusts it and in some years the annual mean is adjusted in the wrong direction (negative hom_ind). In this example the very similar values of the test and the benchmark in the first two years reveal a weak spot of such tool, easily affected by singularities. Aspect that will be treated in the concluding remarks. Finally, when checking the trends (which are presented in the legend), HOM performs better than the other methods, being the only one not over-correcting the trends and being the second closest to the benchmark value.

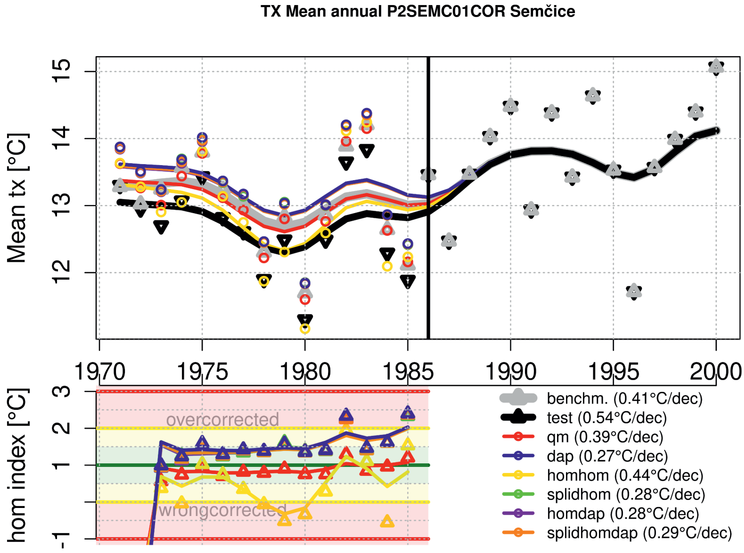


Figure 4.7: *Comparison of homogenized versions of the benchmark series of Semčice, Czech Republic. Color code is set according to table 4.1, black items represents the test series, grey items stand for the value of the benchmark. Top panel shows annual means (points) with running mean (lines), the more the homogenized values get close to the grey line, the better the homogenization has been. This distance is better displayed in the bottom panel, applying the homogenization indicator, explained in equation 4.1, to each yearly value.*

Table 4.2 also indicates that the performance on PD05 is approximately the same among the six methods. At the same time the last six columns, which display the hom_ind applied to the trends, show only a general overestimation of the trends on annual mean of TX. For the annual mean of TN an even distribution around the ideal value (1) is found with the exception of the anomalous value of QM (1.37). The homogenization of the extreme values shows underestimation of almost all the method and all indices. This is observed especially for the 90^{th} percentile of TX with values ranging between 0.42 and 0.63. Furthermore the performance on the trends of the 10^{th} percentile by HOMDAP appears to be very low for both TN and TX. Figure 4.8 shows the distributions that generate these averages. The higher value for QM is due to a distribution that peaks at values larger than 1, indicating a good performance with a slight overestimation. On the other side, the lower averages observed for the other methods are related to a more spread distribution and a higher presence of outliers, which are plotted in single columns out of the vertical solid lines.

A more exhaustive description of the performance of the methods in reproducing the trends on average and extreme indices is shown in figure 4.9. Here the evaluation has been made via scatter plots that allow to compare at the same time the hom_ind results about trends of extreme indices and on the mean values.

Left (Right) panel shows the comparison of hom_ind of trends of annual mean vs 10^{th} (90^{th}) of minimum (maximum) temperatures: all methods show a scattered behaviour. Nevertheless, the lines (contours of half height of the two-dimensional probability density function) help identifying that all the methods homogenize generally well, since they are centred around the black diamond (coordinates (1,1), corresponding to an ideal homogenization on both indices). Thus the anomalous behaviours mentioned above are isolated and not systematic.

In both cases the distribution is more spread along the x axis, indicating a lower performance in the homogenization of extreme values. At the same time there is less ordered behaviour in the contours of the distributions in the case of minimum temperature which is probably due to the more local characteristics of such variable.

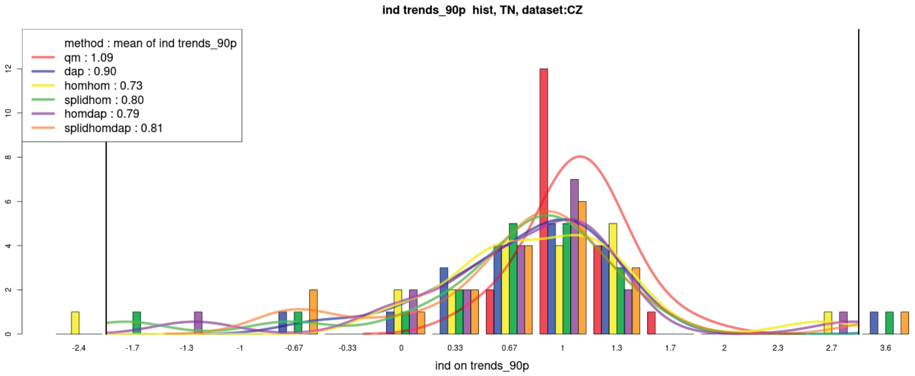
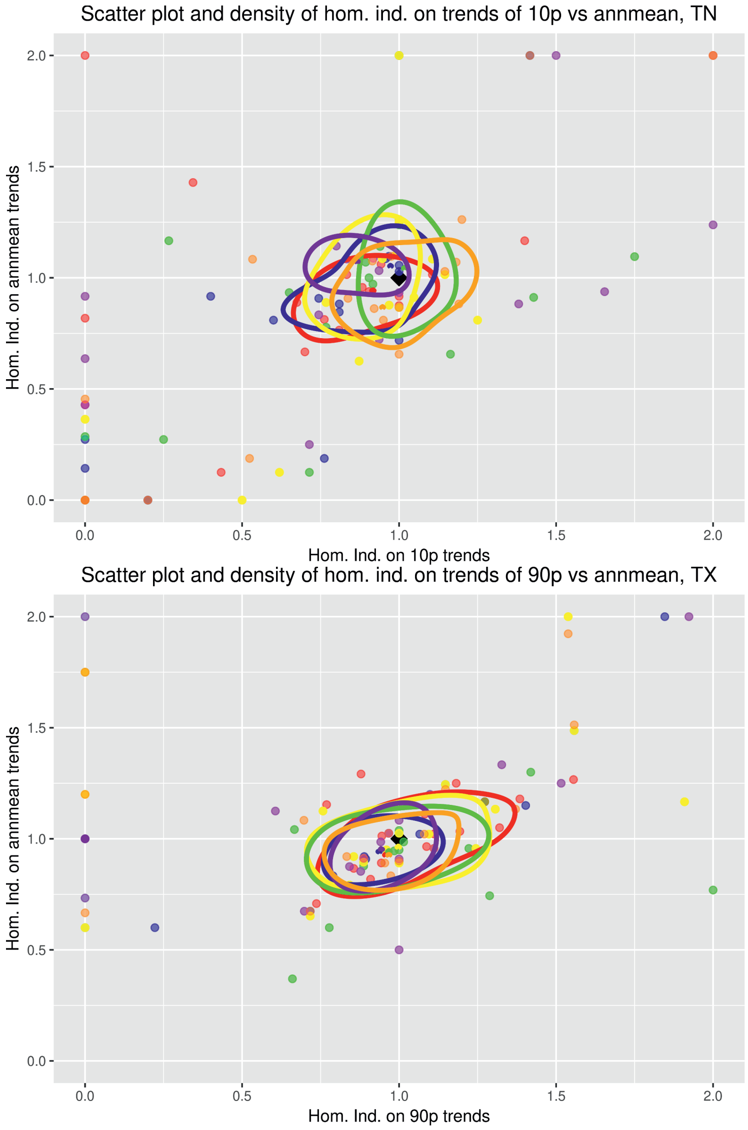


Figure 4.8: *Histograms and density plots of hom_ind applied to the trends of the 90^{th} percentile (90p) of minimum temperature series (TN). Vertical lines define the central area of the histogram, that needs smaller bins in order to better describe the distribution. Values out of such limits are plotted in single columns on the external areas, helping identifying outliers.*



Scatter plot and density of hom. ind. on trends of 10p vs annmean, TN

Scatter plot and density of hom. ind. on trends of 90p vs annmean, TX

Figure 4.9: *Scatter plots with contours of half height of density distributions of homogenization indicator calculated on trends on various indices. Left: TN - hom_ind on trends of 10th percentile vs. annual mean. Right: TX - hom_ind on trends of 90th percentile vs. annual mean. N.b. values which lay out of the graph domain are plotted on the boundaries. Colour code follows table 4.1.*

Table 4.3: *Table reporting the amount of non-adjusted series and non-adjusted data.*

dataset(Euro)	number of series with more than 80% of non-adjusted data		mean of percentage of non-adjusted data [%]	
	TN	TX	TN	TX
Quantile Matching	2	3	17.43	20.73
DAP	3	1	34.22	39.69
HOM	0	0	33.43	27.27
SPLIDHOM	0	0	13.27	17.86
HOM(DAP)	3	3	36.22	40.60
SPLIDHOM(DAP)	4	2	35.18	40.79

4.3.2 European dataset

The scarce geographical density of test and reference series of the European benchmark dataset and the presence of multiple and non-simultaneous breakpoints make that the availability and length of homogeneous sub-series that could be used for the homogenization process are limited. These challenging conditions make that some methods abstain from homogenizing some tests and apply no adjustments on portions of the series. Table 4.3 reports the number of series with a fruitless homogenization (see section 4.2.4 for the definition) indicating an average rate of 1 to 4 failures (out of the 50 test series) for all methods but HOM and SPLIDHOM. These latter two methods adjust a sizeable part of all test series, despite the scarcity of references. Figure 4.10 shows the areas where the methods fail more often: regions with low density (Spain, Italy) and occasionally those on the edges of the denser areas.

Though, while the number of failed series are similar, significant differences are observed on the average of non-adjusted data. The mean of the percentage of non-adjusted data (last two columns of table 4.3) show different behaviours. In the histograms of figure 4.11 QM, SPLIDHOM and HOM present relative narrow distribution, whose low peak values indicate successful performances. Nevertheless QM presents 5 series with 100% of non-adjusted and the high value of the peak of HOM indicates the possibility of issues when homogenizing some of the series. At the same time the other methods show spread distributions indicating high percentage of sparsely non-adjusted data, whose features will be inspected further in the text. This preliminary analysis has allowed to understand how the methods approach the low density of stations, since some methods (as QM) leave the most problematic series unchanged, while others (ProclimBD, HOMDAP and SPLIDHOMDAP) manage to homogenize portions of their daily data.

rRMSE, PD05 and accuracy in the reproduction of the trends have been then evaluated on the homogenized versions, including those series (and those portions of the series) that haven't been adjusted. Table 4.4 highlights an evident lower performance of the

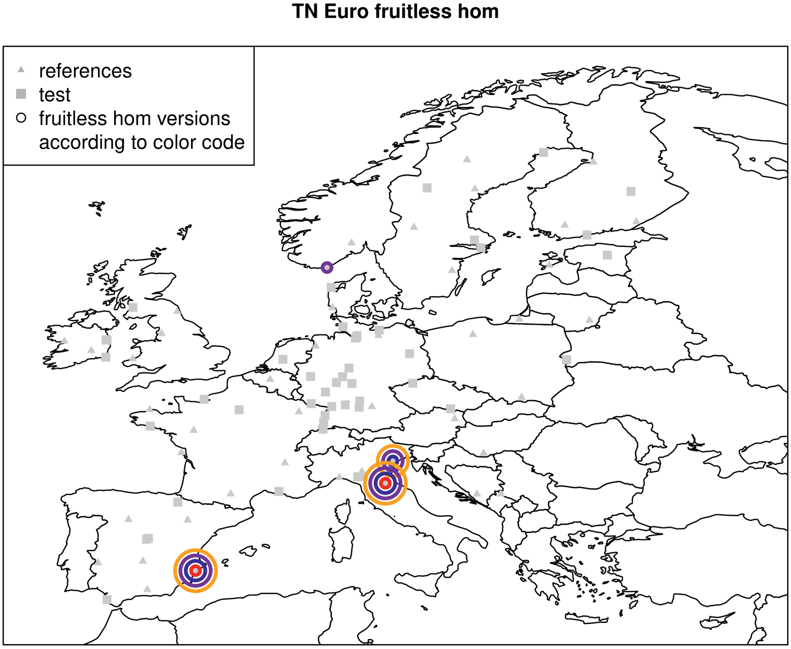


Figure 4.10: *Series whose failing rate is above 80%. Color code follows table 4.1.Sizes of the circles change for graphical purposes and don't represent different magnitudes.*

tested methods compared to the case of the Czech dataset (table 4.2), which can be

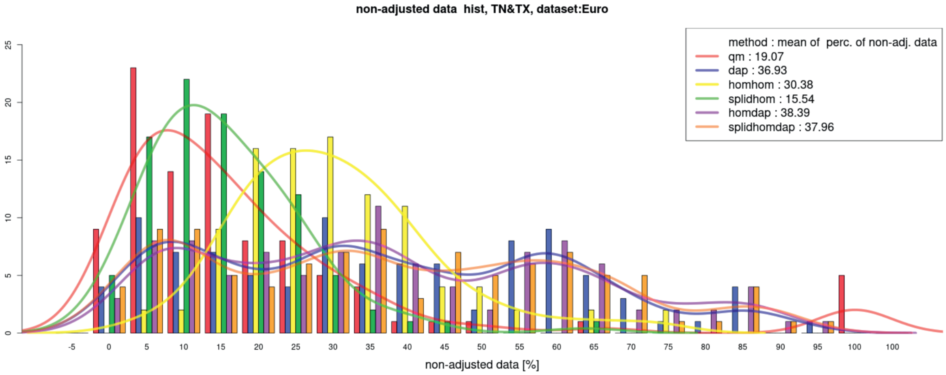


Figure 4.11: *Histograms and density plots representing the percentage of non-adjusted data for each series. Colour code follows table 4.1.*

expected. The reduction of rRMSE after the application of the homogenization methods has considerably a lower magnitude and, at the same time, the PD05 shows a lower increase of percentages. While almost all histograms and pdf show a good improvement (see figure 4.12) compared to the inhomogeneous series, the HOM histogram (yellow) presents a clearly anomalous behaviour, which was anticipated by high averages of table 4.4 and repeats what was already observed in the context of the Czech dataset.

The comparison of the obtained trends with the ones of the corresponding benchmark has revealed a high difficulty in the reproduction of the original trends within the European dataset. While the averages of indicators related to HOM versions are almost always outside of the 0.5-1.5 range (which indicates 50% distance from the benchmark trend), all

Table 4.4: *Results of methods comparison on the European dataset. Relevant negative results are highlighted in bold.*

dataset(Euro)	mean of rRMSE [°C]		mean of PD05[%]		mean of hom_ind of trends of					
					annmean		10p		90p	
	TN	TX	TN	TX	TN	TX	TN	TX	TN	TX
Test Series	1.60	1.28	39.51	46.37						
Quantile Matching	1.25	1.16	43.72	50.41	0.65	0.82	1.01	0.33	0.78	1.23
DAP	1.34	1.21	42.10	48.43	0.55	0.56	0.86	0.91	0.36	1.56
HOM	2.00	1.74	38.20	44.80	1.46	0.31	0.80	1.92	2.00	3.14
SPLIDHOM	1.27	1.21	42.77	47.50	0.87	0.59	0.95	0.79	0.30	1.49
HOM (DAP)	1.42	1.24	40.68	47.85	0.58	0.37	0.63	0.65	0.29	1.63
SPLIDHOM (DAP)	1.35	1.22	41.93	48.51	0.57	0.49	0.69	0.80	0.32	1.51

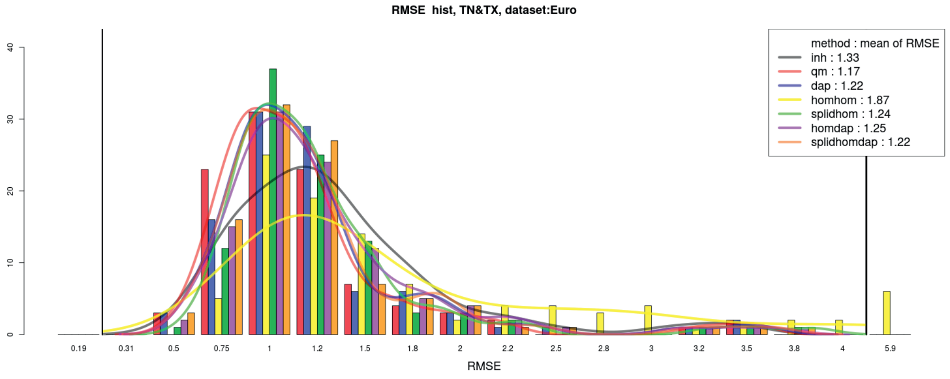


Figure 4.12: *Histograms and density plots of rRMSE (TN & TX together) for the comparison of methods on the European dataset. Lines represent corresponding probability density functions.*

the other methods present similar behaviours, with few exceptions. Worst performances are observed on the trends of the warm extremes (90^{th} percentile). The underestimation of the trends for TN (of all calculated indices) are related to a large presence of very negative results (related to adjustments in the wrong direction) as shown in the top panel of figure 4.13, where a high number of series lie on the left of the boundaries of the histogram. The opposite happens for trends of 90p for TX, where the overestimation is due to a high number of overcorrected series, which lie on the right of the boundaries of the histogram (bottom panel of figure 4.13). In both cases the very high and very low values of the homogenization indicator are related to singularities when the amplitude of the inhomogeneity was very low (i.e. below 0.5 °C). Such a situation implies a low value in the denominator of equation 4.1, thus large absolute value of the indicator.

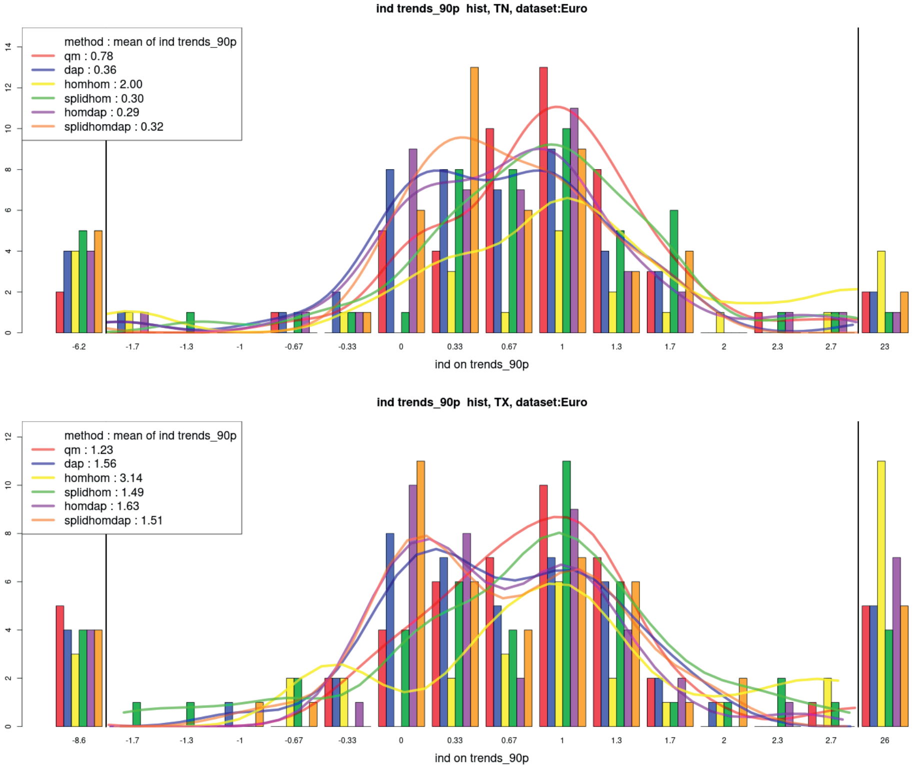


Figure 4.13: *Histograms and density plots of hom_ind calculated on trends of 90p of TN (top) and TX(bottom). Same graphical features as figure 4.8*

An example of such fact is displayed in figure 4.14, where the low magnitude of the breaks is identified but all methods but HOM that over-corrects the first portion of the series (large positive values of the homogenization indicator on yearly 90p of TX). This has a consequence on the calculation of hom_ind on the trend of HOM, which in this case has a value of 6, indicating a correction 6 times bigger than the one needed to reproduce the trend of the benchmark.

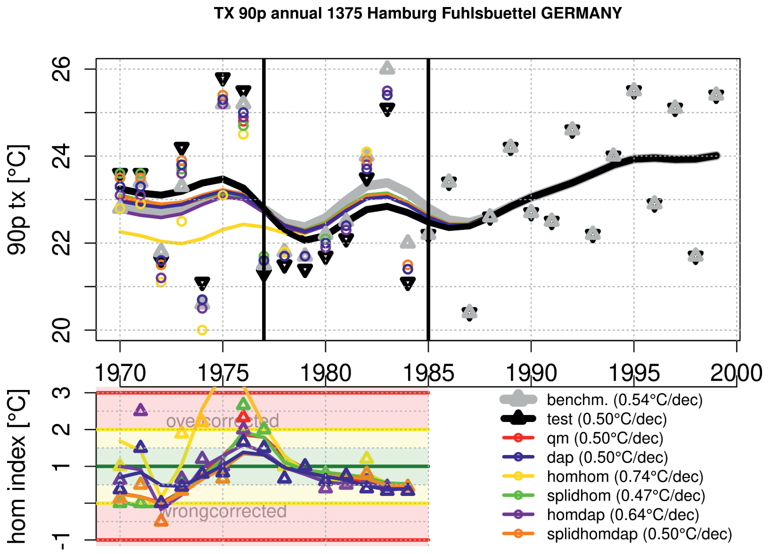
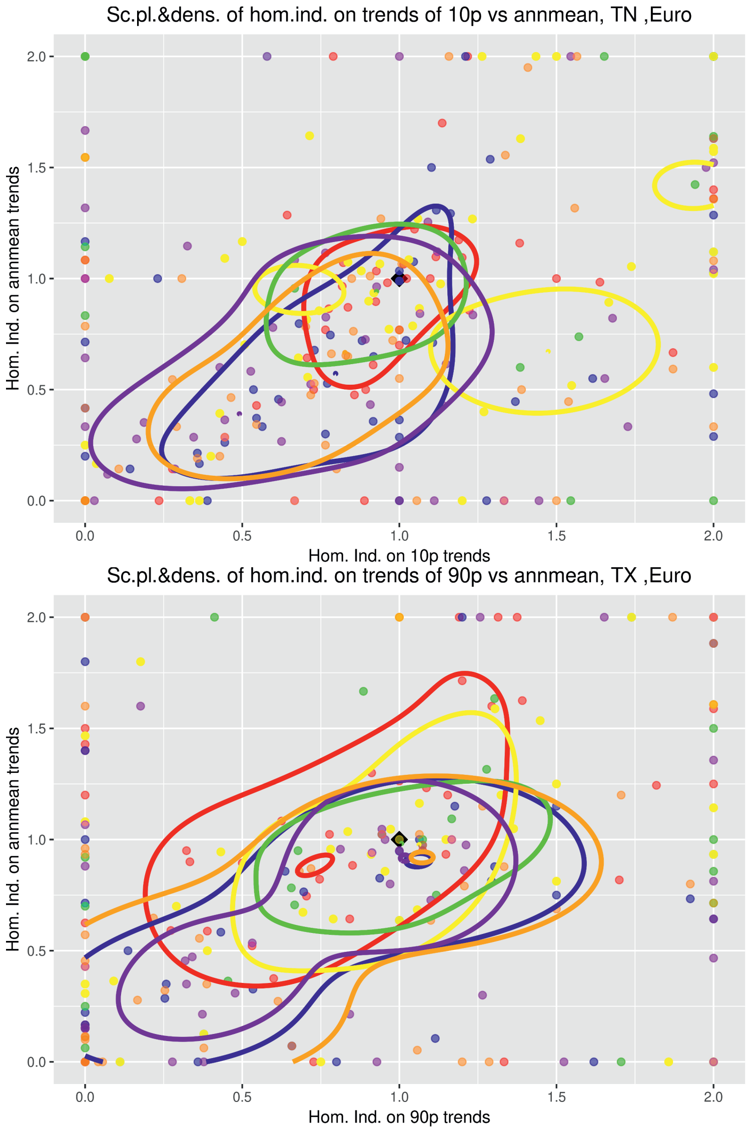


Figure 4.14: *Effects of the homogenization methods on the 90p of TX of Hamburg. Same structure and colour code as figure 4.7.* N.b. Good accordance among the results imply that some of the lines are overlapped

Special attention has to be given to the particular shape of the distribution of DAP, HOMDAP and SPLIDHOMDAP in the bottom panel of figure 4.13. This is related to the conservative philosophy of DAP, which is to preserve original data rather than risking wrong adjustments. Consequence of this choice is the high percentage of non-adjusted data observed in figure 4.11. The peak close to 0, which is related to trends which are not altered during the homogenization, indicates a high number of non-changed data in the warmest records of these series. The same behaviour is found in the scatter-plots of the right panel of figure 4.15, where these three methods present a bimodal structure, with one of the two maxima that tends to get close to the (0,0) point, which represents uncorrected data. This scatter-plot, together with the one related to TN temperatures (left panel), displays a very scattered distribution as well as broader and less defined densities, reflecting the difficulty in homogenizing such dataset. Nevertheless QM, and especially SPLIDHOM, appear to perform well, moving the peak in the distribution near (1,1).



Sc.pl.&dens. of hom.ind. on trends of 10p vs annmean, TN ,Euro

Sc.pl.&dens. of hom.ind. on trends of 90p vs annmean, TX ,Euro

Figure 4.15: *Same as figure 4.9 but for the European dataset.*

4.4 Discussion

The higher density and correlations of the Czech dataset (see table 4.5) makes that almost all the methods perform well, homogenizing the whole series with similar results.

Table 4.5: *Average distance and correlation in the two datasets. Please note that the Czech dataset has an homogeneous distribution of the stations, while the density of stations in the European varies considerably among the regions.*

	Czech		European	
	TN	TX	TN	TX
Average distance [km]	40.4	38.3	198.0	220.1
Average correlation	0.896	0.928	0.629	0.615

Nevertheless, even under favourable conditions a method like HOM has been affected by statistical issues. The causes of this might be inspected in the software of this method, nevertheless such research is not in the scope of the present work.

On the other side, the comparison of the methods on a more challenging dataset has helped to test the robustness of the procedures and has revealed different behaviors. The evaluation of the percentages of unchanged data has highlighted how approaches in the application of the adjustments vary among the considered methods. Those that use DAP settings (DAP, HOMDAP and SPLIDHOMDAP) require an improvement in the candidate-reference correlation (at least 0.05) in order to allow the adjustment of a daily datum, preferring to leave the data unchanged when such constraint is not fulfilled. This is evident in figures 4.11,4.13 and 4.15 and highlights how procedural choices lead to similar results even when statistical properties (see section 4.1.1) of the methods are considerably different. The aforementioned choices are the effect of a more conservative strategy that can be adopted by homogenizers. This aims at giving (when in doubt) priority to original data over homogenized ones, which can be required for particular studies or projects.

The spread distribution of percentage of non-adjusted data in the series observed in figure 4.11 is the most evident effect of the conservative constraints. Though, this is considerably different from the distribution observed for example in the case of quantile matching. This last method implements its conservative strategy with a different approach. It evaluates the possibility of applying adjustments considering the whole candidate and reference series: when correlation is too low or the overlapping period is too short, the whole series is left unchanged, explaining the peak of unchanged data at 100%.

Finally, while the non-adjustment criteria of QM affect the whole distribution of daily data, figures 4.13 and 4.15 have shown that the unchanged data related to the methods with DAP parameters are more concentrated around the warm extremes. This is probably due to a lower correlation gain when homogenizing summer extremes.

4.5 Summary and Conclusions

A comparison between homogenization methods (Quantile Matching [Trewin, 2013; Squintu et al., 2019], DAP [Štěpánek et al., 2009], HOM [Della-Marta and Wanner, 2006] and SPLIDHOM [Mestre et al., 2011]) has been performed on two benchmark datasets. The two dataset (Czech Republic, Europe) are built with analogous methods, replacing data of homogeneous series (*benchmarks*) with data coming from homogeneous segments of neighbouring series, thus generating non-homogeneous series to be adjusted (*tests*). While the Czech dataset has high quality and station density, in the European dataset stations are sparse and homogeneous reference series represent a minority (44% of the set, down to 20% in some areas). The performance of the homogenized versions (*candidates*) has been evaluated using metrics such as the rRMSE (RMSE of the difference series between candidate and corresponding benchmark), PD05 (percentage of values in the difference series below 0.5°C). These metrics have revealed a good performance of all the methods but HOM, which presents a high number of outliers in both datasets. Trends on indices such the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentile (together with the annual mean) have then been tested. This has allowed to test performance of the homogenization methods in reproducing extreme values and related indices. Such comparison has been operated via the introduction of a homogenization indicator (hom_ind) that can be applied to any indices or measured values. This determines the goodness of the homogenization on a clear common scale: wrong adjustment (< 0), no adjustment (0), partial adjustment (between 0 and 1), ideal adjustment (1) and overcorrection (> 1). Plotting results of hom_ind helps identifying anomalous behaviours and systematic biases. In particular it has been observed that, while in the Czech dataset all methods have performed well with minor exceptions, the more challenging European dataset has revealed several criticisms, among all in the homogenization of warm extremes for three methods (DAP,HOMDAP,SPLIDHOMDAP).

Though, the structure of the equation that defines hom_ind has shown to be vulnerable in case of low difference between test and benchmark values. These situations can lead to singularities (i.e. very large values of hom_ind) which in reality are related to reasonable values. At the same time conditions of no difference between test and benchmark are affected by a very low signal-to-noise ration. This means that the homogenization process presents high values of uncertainty, making the result hard to be defined in terms of quality. Thus a more sophisticated version of the equation (4.1) may be developed in order to avoid this issue, since inhomogeneities with low magnitude are important challenges of benchmark (and real) datasets and any homogenization method needs to be able to identify and account for them. Further changes will also be needed in the technique for the construction of benchmark datasets, for example widening the size of the datasets, extending the covered time interval, inserting plateau breaks (as suggested by Domonkos and Coll [2017]) or giving more relevance to mountain areas where a minimal change of altitude may bring to large inhomogeneities. Nevertheless the use of two different benchmark datasets, as done here, has shown to be a powerful tool in order to perform a complete analysis of the capabilities of the methods.

In this study the methods are compared in their original version, without further elements but those described in their original papers. Though it is clear that the different approach in the selection of reference series affects the performance of the methods themselves. In particular the use of a single or a set of references (real record or average of surrounding series) make the methods respond differently to the statistical features of the references. For this reason a further study would be needed with the aim of comparing the homogenization methods in similar conditions of reference selection, analyzing how they would behave in all the possible situations (i.e. set of references, average of references or single reference). Nevertheless this work has allowed a comparison under different aspects of the selected methods and has given clear indications about their performances.

It's fundamental to stress that the selection of the best homogenization method is not in the goals of this work: such choice depends on the needs of each individual user. Studies like this aim at stimulating the debate on the field of homogenization of temperatures, stressing criticisms of some methods and indicating to future programmers which aspects (e.g. management of low density of stations, importance of extreme values) should be better developed to reach an always higher quality of homogenization on a daily basis. For example particular care should be given in managing the adjustment of extreme values which have been proven to need a different treatment with respect to average values. On the other hand homogenization methods must be able to account for the most various geographical patterns. This implies that the procedure is required to be flexible enough to adapt to conditions of areas with very high and very low data quality and correlation between series. At the same time, users should choose the most suitable homogenization method evaluating these aspects, considering especially the station density and the statistical properties of the climate of the interested area (i.e. mountain, plain, coastal climates, etc.). For these reasons the belief of the benefit [Venema et al., 2013] that publication of studies about the comparison of homogenization methods have on the field has become stronger.

Finally a short consideration is needed about the specific case of the European Climate Assessment and Dataset. This includes areas with extreme high density (Germany, Sweden, etc.) and areas with low density (Northern Africa, Middle East, parts of Mediterranean Europe, etc.), thus a method that is able to face both data scarce as data rich conditions is needed. Quantile Matching was applied earlier to the whole ECA&D dataset [Squintu et al., 2019] and the good performance of this method on both the Czech dataset and the European dataset, characterizing data-rich and data-scarce conditions respectively, and the similarity in the results to a widely known and recognized method as SPLIDHOM, gives confidence in the homogenization of ECA&D using Quantile Matching.

Acknowledgements

We acknowledge the data providers in the ECA&D project (www.ecad.eu). Funding has been received from the Copernicus Climate Change Service through contract C3S_311a_Lot4.

Chapter 5

Evaluation of trends in extreme temperatures simulated by HighResMIP models across Europe

Abstract

Simulation of the climate of the past day by day is an important tool for the validation of of climate models. The comparison with observed daily values allows to assess their reliability and the soundness of their projections on the climate of the future. Frequency and amplitude of extreme events are fundamental aspects that climate simulations need to reproduce. These have high impacts on economy and society. The ability of simulating them will allow policy makers in taking better measures to face the climate changes. In this work six models developed within the High Resolution Model Intercomparison Project are compared over Europe with the homogenized version of the E-OBS gridded dataset. This is done first comparing averages, extremes and trends of the simulated summer maximum temperature and winter minimum temperatures with the observed ones. Extreme values have been analyzed making use of indices based on the exceedances of percentile-based thresholds. Winter minimum temperatures are generally underestimated by models in their averages (down to -4°C of difference over Italy and Norway) while simulated trends in averages and extreme values are found to be too warm on western Europe and too cold on eastern Europe (e.g. up to a difference of -4% per decade on the number of Cold Nights over Spain). On the other side the models tend to underestimate summer maximum temperatures averages in Northern Europe and overestimate them in the Mediterranean areas (up to +5°C over the Balkans). The simulated trends are too

warm on the North West part and too cold on the South East part of Europe (down to -3%/dec. on the number of Warm Days over Italy and Western B alkans). These results corroborate the findings of previous studies [Min et al., 2013; Sillmann et al., 2013] about the underestimation of the warming trends of summer temperatures in Southern Europe, where these are more intense and have more impacts. A comparison of the high resolution models with the corresponding version in CMIP5 has been performed comparing the absolute biases of extreme values trends. This has shown a slight improvement for the simulation of winter minimum temperatures, while no signs of significant progresses have been found for summer maximum temperatures.

This chapter was originally submitted and is currently under revision as:

Squintu, A. A., van der Schrier, G., van den Besselaar, E., van der Linden, E., Scoccimarro, E., Roberts, C., and Klein Tank, A. (2020b). Evaluation of trends in extreme temperatures simulated by highresmip models across europe (submitted). *Climate Dynamics*

5.1 Introduction

5.1.1 Climate simulations

A change in the frequency and intensity of climatic heat extremes has important impacts of the society and the economy, as most of these events have serious effects on agriculture, energy demand, transportation industry, health, etc. For this reason a prediction of future climate that realistically assesses changes in the evolution of extreme events is fundamental in order to understand the challenges that need to be faced. The climate models used for the production of future projections are also used to simulate the climate of the recent past. This is done by taking as input verified historical observed boundary conditions (e.g. land use) and using observed values as internal (artificial and natural, such greenhouse gases and aerosol concentration, volcanic eruptions) and external forcings (e.g. solar irradiance).

The simulations of the climate of the past are an important challenge for climate models and, since the statistics of these simulations can be compared with observed values, are crucial for the validation of projections to the future [Flato et al., 2014].

The Climate Model Intercomparison Project 3 (CMIP3) [Meehl et al., 2007] and the CMIP5 [Taylor et al., 2012] have contributed in collecting and comparing all the available climate simulations and projections. CMIP5 uses an improved specification of historical forcings and contains experiments with longer temporal perspectives than its predecessors [Taylor et al., 2012; Flato et al., 2014], and it has shown a clear improvement in model performance on temperature simulations compared to CMIP3 [Flato et al., 2014]. More recently the project PRIMAVERA (funded within the European Union Horizon 2020 project) has worked in the frame of the CMIP6 HighResMIP protocol . Here a coordinated set of experiments has been designed to assess both standard and enhanced horizontal-resolution simulations in the atmosphere and ocean (with up to 0.25° in the atmosphere) [Haarsma et al., 2016].

5.1.2 Gridded datasets as comparison-term for climate simulations

The comparison of these climate simulations against observations, focusing e.g. on long term trends, is a powerful tool for the evaluation of how the models reproduce the climate under observed conditions and forcings.

Such validation has important implications in determining how reliable the same models are when applied to the future decades and what they should improve in order to produce better projections [Bhend and Whetton, 2013].

The evaluation of climate simulations has been the subject of several studies and has been improved in the last years [Flato et al., 2014]. These have focused on the intercomparison of the models or on the comparison with reanalyses, observed individual series or, more recently, gridded observational datasets. The choice of the used reference is fundamental and needs to be performed with care [Sillmann et al., 2013].

The use of observations is preferable [Gleckler et al., 2008; Flato et al., 2014] since reanalyses, even though they can be used as a replacement for areas with sparse observational coverage, are affected by a poor representation of extremes.

The use of individual station data is a very direct approach, but hampered because the gridbox value of a model represents an area-average whereas the station observation is a point-value. This is particularly problematic for climatic extremes.

For these reasons the validation against observational gridded datasets is more appropriate. Their grid squares represent area-averages as in the case of the models. In addition, such datasets provide a homogeneous spatial distribution and remove some observational noise, which is a typical issue of station data [Cornes and Jones, 2013]. Nevertheless such products, especially with a daily resolution, have been lacking for long time [Kiktev et al., 2003; Kharin et al., 2005] and the possible comparisons are still limited in space and time [Flato et al., 2014].

5.1.3 Simulation of extreme events

The E-OBS gridded dataset [Haylock et al., 2008; Cornes et al., 2018], used in this study, matches the high spatial and temporal detail of this new generation models, see section 5.2.2 for more details.

The works of [Kharin et al., 2005] first and, more recently, [Sillmann et al., 2013] have shown that climate simulations reproduce time averaged values, like monthly means, better than the short duration extreme events. Though these have great impact on economy and society and, when compared to average values, present different trends and behaviors [Hartmann et al., 2013]. It is widely accepted that changes in the climate are driven by variations of the average [Scherrer et al., 2005]. Nevertheless debate has taken place on the magnitude of the change of variability [Alexander et al., 2006; Simolo et al., 2010; Morak et al., 2011; Donat and Alexander, 2012], since extreme events depend more on this than on the average [Katz and Brown, 1992]. Changes in variability are shown to lead to amplified effects on values exceeding a certain threshold [Della-Marta et al., 2007] and a change in the mean insufficiently explains particular record-breaking events, such as the heatwave of 2003 [Schär et al., 2004]. This has motivated researchers to go beyond the validation of the model average European climate and to focus on the most impacting aspects [Stainforth et al., 2013].

While the analysis of changes in aggregated averages has been the subject of several studies [Gleckler et al., 2008; van Oldenborgh et al., 2009], the attention on the assesment of variability in extreme events has increased in the last years [Kiktev et al., 2003; Kharin et al., 2005, 2007; Sterl et al., 2008; Min et al., 2013; Sillmann et al., 2013, 2014b,a]. A few studies have inspected the simulation of events with a return time of the order of years, the so-called hard extremes [Kharin et al., 2005, 2007; Sterl et al., 2008]. Calculations of trends on such statistics often lacks significance [Frich et al., 2002] which motivates the use of indices related to soft (or moderate) extremes that occur one or multiple times every year [Klein Tank and Können, 2003]. Indices related to soft extremes have been established by the Expert Team on Climate Change Detection and Indices (ETCCDI) [ETCCDI, 2009], following previous works [Frich et al., 2002]. One kind of soft extremes indices is based on the most extreme yearly value of the records, e.g. TXx (warmest maximum temperature of the year) or TNn (coldest minimum temperature of the year). Such indices present a high interannual variability, which makes it difficult to calculate trends on relatively short periods. Furthermore, these indices are very sensitive to quality issues in the observational data, especially if the quality issues occur at a station in a data sparse area.

Indices based on absolute or percentile-based thresholds are good alternatives to focus on climatic extremes. Absolute thresholds cannot be applied with same efficacy to all climates, as in the case of Frost Days (number of days with minimum temperature, TN, below 0°C) to warm areas (e.g. southern Mediterranean or Middle East) or Summer Days (number of days with maximum temperature, TX, above 25°C) to high-latitude or high-altitude regions (e.g. Iceland or northern Scandinavia). Such features make these indices hard to use for geographical comparison. In addition, if a model has a bias in the mean values, this is likely to have systematic effects on the indices based on fixed threshold. Such issues do not apply to indices with percentile-based thresholds, for example: TN10p, percentage of days with daily minimum temperature below the 10th percentile and TX90p, percentage of days with daily maximum temperature above the 90th percentile. Both percentile thresholds are calculated considering the data of the series in a long reference period, usually thirty years. These are site specific, since the thresholds change for each grid point according to the local climatic features. Thus they are immune to biases in mean values and applicable to any climate [Klein Tank and Können, 2003; Kiktev et al., 2003; ETCCDI, 2009; Sillmann et al., 2014b].

The 5th International Panel on Climate Change (IPCC) assessment report, while indicating progresses for CMIP5 over CMIP3 in the bias of the models (with relevant discrepancies in Northern Europe) [Flato et al., 2014; Bhend and Whetton, 2013], criticized the description of the seasonal cycle. This showed a slight underestimation over the Mediterranean and overestimation over Eastern Europe [Flato et al., 2014]. At the same time the simulation of extreme events in the second half of the 20th century has improved [Flato et al., 2014; Sillmann et al., 2013]. Nevertheless, in the case of regional models forced with reanalysis data, Min et al. [2013] reports an underestimation in trends of extreme warm temperature values in the European and Mediterranean region. In particular, none of the models considered by Min et al. [2013] identified a significant trend while observational datasets showed significant changes.

5.1.4 Purpose of the paper

Since Europe belongs to one of the most sensitive areas to climate change [Giorgi, 2006; Brown et al., 2008] the current study takes advantage of the new homogenized gridded dataset over Europe and aims to validate global high resolution models that have been made available in the PRIMAVERA project. This includes an assessment of the improvement of these high resolution models in reproducing observed trends in comparison to their lower resolution version, following other works that have evaluated the progresses of HighResMIP in the analysis of other phenomena, such as tropical cyclones [Roberts et al., 2020].

5.2 Data & Methods

5.2.1 Models

The tested simulations have been developed in the frame of the PRIMAVERA Project, that aims at increasing spatial resolution of climate models. Six models have been analyzed in latest high resolution (HR) version and in a previously existing lower resolution (from now on referred to as lower resolution, LR), focusing on the period from 1970-2014 (see section 5.2.2) and considering the region enclosed between 22°W and 50°E and 20°N and 76°N. The variables considered in this study are minimum temperature (TN) and maximum temperatures (TX) on a daily resolution. Each model taking part in PRIMAVERA has contributed with several experiments, the one that has been used for this work is named "highres-SST-present". This consists of an atmosphere-only integration, forced with observed SST, observed sea-ice concentration and external radiative forcings over the period 1950-2014. Each model has a different spatial resolution and a different number of ensemble members. Tables 5.1 (HR) and 5.2 (LR) summarize the characteristics of the used models and the availability of the ensemble members as of 23$^{\mathrm{rd}}$ of September, 2019.

The *ECMWF* model has native resolution Tco399 ($\sim$ 25 km) for HR and Tco199 ($\sim$ 50 km) for LR. In the frame of PRIMAVERA they have been provided in a regridded version, respectively to 0.25° and 0.5° constant latitude-longitude regular grids, more details in [Roberts et al., 2018]. The *Ec-Earth3P* model runs at the resolution TL511 for HR and TL255 for LR on a non-regular latitude-longitude grid. The scripts used for the indices calculation (5.2.3) require regular grids, making it necessary to regrid the Ec-Earth3P model.

Table 5.1: Characteristics of the high resolution models that have been used. Resolutions have been rounded at the second decimal digit. Insitute acronyms and corresponding references for the models: Centro euro-Mediterraneo sui Cambiamenti Climatici (CMCC), [Cherchi et al., 2019]; Centre National de Recherches Météorologiques - Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CNRM-CERFACS), [Voldoire et al., 2019]; EC-Earth, [Haarsma, 2020]; European Centre for Medium-Range Weather Forecasts (ECMWF), [Roberts et al., 2018]; MetOffice - Hadley Centre (MOHC), [Roberts et al., 2019]; Max-Planck-Institut für Meteorologie (MPI-M), [Gutjahr et al., 2019].

Acronym	**CMCC**	**CNRM**	**Ec-Earth**	**ECMWF**	**HadGEM3**	**MPI**
Model name	CMCC-CM2-VHR4	CNRM-CM6-1-HR	EC-Earth3P -HR	ECMWF-IFS-HR	HadGEM3-GC31-HM	MPI-ESM1-2-XR
Institute	CMCC	CNRM-CERFACS	EC-Earth-Consortium	ECMWF	MOHC	MPI-M
# of ens. memb.	1	1	3	6	3	1
ens. memb. labels	r1i1p1f1	r1i1p1f2	r1i1p1f1 r2i1p1f1 r3i1p1f1	r1i1p1f1 r2i1p1f1 r3i1p1f1 r4i1p1f1 r5i1p1f1 r6i1p1f1	r1i1p1f1 r1i2p1f1 r1i3p1f1	r1i1p1f1
Long. res. [°]	0.31	0.5	0.49 (regridded)	0.5	0.35	0.47
Lat. res. [°]	0.23	0.5	0.35 (regridded)	0.5	0.23	0.47

Table 5.2: Characteristics of the low resolution models that have been used. Resolutions have been rounded at the second decimal digit.

Acronym	**CMCC lowres**	**CNRM lowres**	**Ec-Earth3 lowres**	**ECMWF lowres**	**HadGEM3 lowres**	**MPI lowres**
Model name	CMCC-CM2-HR4	CNRM-CM6-1	EC-Earth3P	ECMWF-IFS-LR	HadGEM3-GC31-LM	MPI-ESM1-2-HR
Institute	CMCC	CNRM-CERFACS	EC-Earth-Consortium	ECMWF	MOHC	MPI-M
# of ens. memb.	1	1	2	8	5	1
Ens. memb. labels	r1i1p1f1	r1i1p1f2	r1i1p1f1 r3i1p1f1	r1i1p1f1 r2i1p1f1 r3i1p1f1 r4i1p1f1 r5i1p1f1 r6i1p1f1 r7i1p1f1 r8i1p1f1	r1i1p1f1 r1i2p1f1 r1i3p1f1 r1i14p1f1 r1i15p1f1	r1i1p1f1
Long. res. [°]	1.25	1.4	0.98 (regridded)	1	1.88	0.94
Lat. res. [°]	0.94	1.4	0.7 (regridded)	1	1.25	0.94

5.2.2 E-OBS

The reference used for the evaluation of the models is the E-OBS for TN and TX [Haylock et al., 2008; Cornes et al., 2018]. It comes as a 100-members ensemble, whose spread increases in areas with low station density, indicating a larger uncertainty. In this work only the ensemble mean is considered. E-OBS is based on the station data of the European Climate Assessment & Dataset (ECA&D) [Klein Tank et al., 2002], which collects data of thirteen variables from more than 19000 stations located in all countries of the European and Mediterranean region. Almost 10000 of these stations provide temperature data. These are provided by National Meteorological and Hydrological Services, universities or private companies and range from late 18th century to current times. However, relocation of stations, instrumentation changes and variations in the surroundings of the meteorological stations affect the quality of ECA&D temperature temporal series related to such stations (and therefore E-OBS), reducing the reliability for temporal analyses. For this analysis, a modified version of E-OBS is constructed based on recent work on the homogenization of the temperature series of ECA&D [Squintu et al., 2019, 2020c]. This work has removed a large part of the inhomogeneities and makes it possible to smoothly combine series that belong to neighbouring stations, gathering data into one long-running homogeneous series, called blended series. This considerably improves the input data for E-OBS. A data set of long and homogenized series are the prerequisite to assess climatic change [ETCCDI, 2009; Jones and Wigley, 2010].

For the purpose of this work, only the blended series that start before 1970 and that stop after 2014 have been considered in the construction of a special version of E-OBS, called E-OBS.hom. This selection aims to have a constant number of blended series contributing to each grid-point, avoiding that the changes in station density might introduce inhomogeneities. Table 5.3 indicates that there is not a drastic change in the number of blended series choosing 1970 or 1980 as starting point, thus 1970 has been chosen in order to work with a longer period.

Table 5.3: Number of series which are continuously active during the indicated time interval. Percentage is calculated with respect to the total amount of blended homogenized series (first row). As expected, longer periods have lower amount of series which are active during the whole interval.

period	# of TN series (*% over the total*)	# of TX series (*% over the total*)
total: no time restriction	3110 (*100%*)	3086 (*100%*)
1950-2014	612 (*19.7%*)	603 (*19.5%*)
1960-2014	997 (*32.1%*)	983 (*31.8%*)
1970-2014	1268 (*40.7%*)	1248 (*40.4%*)
1980-2014	1376 (*44.2%*)	1353 (*43.8%*)

5.2.3 Data analysis

In this work the minimum temperatures in the winter months (*DJF*) and the maximum temperatures in the summer months (*JJA*) have been analyzed. After checking the bias of the seasonal averages (TNavg-DJF and TXavg-JJA), for each grid-point the indices TN10p and TX90p have been calculated on a seasonal level, obtaining respectively TN10p-DJF and TX90p-JJA. In both cases the percentile thresholds have been calculated over the 1981-2010 period, making use of the bootstrapping approach introduced by [Zhang et al., 2005].

In order to perform a grid-point by grid-point comparison the E-OBS indices have been regridded with a bilinear procedure to the native grid of each model (with the exception of the "substitute" grid used for EC-Earth3, see table 5.1), creating six versions of remapped E-OBS for each index.

At this point for each dataset and each grid-point the trends on the indices on the 1970-

2014 period have been obtained. [1] Calculation of trends has been done following the Sen's slope method [Sen, 1968], which is more robust than a least square approach and does not require the assumption of a normal distribution [Sen, 1968; Alexander et al., 2006; ETCCDI, 2009]. Some model experiments have been run in ensemble mode and in order to obtain the ensemble means, for each model the trends on each grid-point related to the ensemble members have been averaged. Each ensemble mean has been compared to the corresponding E-OBS regridded dataset taking the difference of the trends on each grid-point. The difference has been considered significant when the 95% interval of each trend on E-OBS and the 95% interval of each corresponding trend on the model don't overlap. This process, applied to both high and low resolution versions, has allowed to detect areas in which the models underestimate or overestimate the trends that have been seen in observational datasets. Finally, the absolute trend bias is defined as the absolute difference between the trend in the model and the trend in the E-OBS dataset. This bias has been calculated using the HR and the LR models.

To compare the HR absolute trend bias with its corresponding LR one, a new temporary dataset has been created on the grid resolution of the HR model (LRtoHR). These LRtoHR grid-points have been filled by using the absolute trend bias of the LR grid-point that overlaps with the LRtoHR grid-point. This is done in order to better inspect the local impact of the increasing resolution, which would be lost in case of a comparison with regridding to lower resolution.

The HR and the LRtoHR absolute trend biases have been compared by taking the difference as shown in equation 5.1 .

$$abstrendbias_{HR} - abstrendbias_{LR} = |trend_{HR} - trend_{E-OBS}| - |trend_{LR} - trend_{E-OBS}| \tag{5.1}$$

If this metric produces a negative value, then the HR absolute trend bias is lower than the one of LR, thus the trend for HR is closer to the observed one, indicating a better performance. On the other side a positive value indicates that the HR's performance is worse than the corresponding LR. The aim of using absolute trend biases is to assess if the HR trends are closer to the E-OBS trends than corresponding LR trends, independently from the sign of the trend difference. If the comparison was performed with trend biases, it would have only communicated if the HR models simulate warmer (positive result) or

[1] Note that the indices for extreme values used in this work are site specific. This means that, by definition, for each gridpoint of each model the percentage of days in the 1981-2010 period above (below) the the 90th (10th) percentile is exactly 10 %. Therefore it is not expected to observe a significant difference between the percentage of values exceeding the thresholds in the models and in the observations. These, if present, would be only related to the larger considered period (1970-2014) and would not carry significant meaning. For this reason the analysis has been conducted directly on the trends, which describe the changes in the distribution shape.

colder (negative result) trends than the corresponding LR models, see equation 5.2.

$$\begin{aligned} trendbias_{HR} - trendbias_{LR} &= (trend_{HR} - trend_{E-OBS}) - (trend_{LR} - trend_{E-OBS}) \\ &= trend_{HR} - trend_{LR} \end{aligned} \tag{5.2}$$

5.3 Results

5.3.1 Minimum temperatures

Bias in winter averages

The considered HR models show strong differences in the reproduction of TNavg-DJF, see Figure 5.1. The largest mean bias is found for CMCC (+2.96 °C), while Ec-Earth3, ECMWF and HadGEM3 underestimate on average the minimum temperatures, with a common exception on Northern Scandinavia, far from the coast. At the same time an underestimation of winter TN over Italy and Norway is found in 5 models. MPI and CNRM perform best in terms of mean biases and present considerably lower extension of the shaded area. These are present when the simulated TNavg-DJF is significantly different from the observed one (i.e. absence of overlap between the 95% confidence interval of the two terms of the difference).

Trends in winter averages

Trends on the TNavg-DJF of the models in the 1970-2014 period are compared against the same indices of E-OBS. All models reproduce very well the trends in winter TN. The mean trend biases, Figure 5.2, range between -0.16 °C per decade (°C/dec) (CNRM) and +0.02 °C/dec (ECMWF). This indicates a tendency in simulating lower trends over the continent, but mainly for Eastern Europe. Nevertheless recurring positive biases are found over the Kola Peninsula (NorthWestern Russia, 6 models out of 6), in the Balkans (5/6) and along the European coast of the Western Mediterranean (4/6).

Trends on cold extremes

Trends in winter cold extremes as TN10p-DJF are more challenging and Figure 5.3 shows these trends for the HR version of the models. While HadGEM3 (mean trend bias: -1.25 %/dec) and, less strongly, ECMWF (mean trend bias: -0.65 %/dec) simulates a lower trend of number of days below the 10^{th} percentile almost over the whole continent, thus having warmer trends than observed, CMCC presents a large contrast between Western and Eastern Europe. In particular a wide area over Iberia, Southern France and the Alps is found where the differences are significant and exceed -4 %/dec, indicating a poor representation of the trends in these areas. The overestimation of the trends (colder trends than what is observed) over Eastern Europe is common to 4 models, while warmer trends over the Mediterranean area are simulated by all models but MPI. This last one is the only model whose mean bias is not significantly negative and does not present pronounced patterns, with the exception of having a too strong warming in TN10p-DJF over Sweden and Norway, in common with three other models. The combination of the tendency to warmer simulated trends in TN10p-DJF together with the fair representation of trends in average values, imply that these HR models have simulated a winter climate with similar average characteristics but fewer cold events, symptoms of a narrowing daily minimum temperature distribution.

The patterns in the trend bias in the HR models can be compared to the low resolution models, whose results are shown in Appendix 5.A. Figure 5.9 shows that the LR models present similar patterns in the trend biases as HR models s.

Figure 5.4 presents the difference in absolute trends biases of TN10p-DJF between HR and LR, see section 5.2.3. Negative values (green) indicate that HR has lower absolute trend bias than LR for that specific grid-point, thus it is performing better. Only CMCC clearly shows an area with worse absolute trend biases over Central Europe (where very large trends are simulated), which contrasts with the strong performance of the same model over Eastern Europe. Despite of this, the mean absolute trend biases over the whole continent are reduced for almost all the models, indicating a general improvement in the description of the cold extremes between low and high resolution. The best improvement is found for HadGEM3 (-0.51%/dec), while the only worsening, out of the considered models, is for ECMWF (+0.17 %/dec) whose LR version is found to perform the best among the others, see 5.A. The model with the lowest mean absolute trend bias in high resolution is MPI (0.61 %/dec).

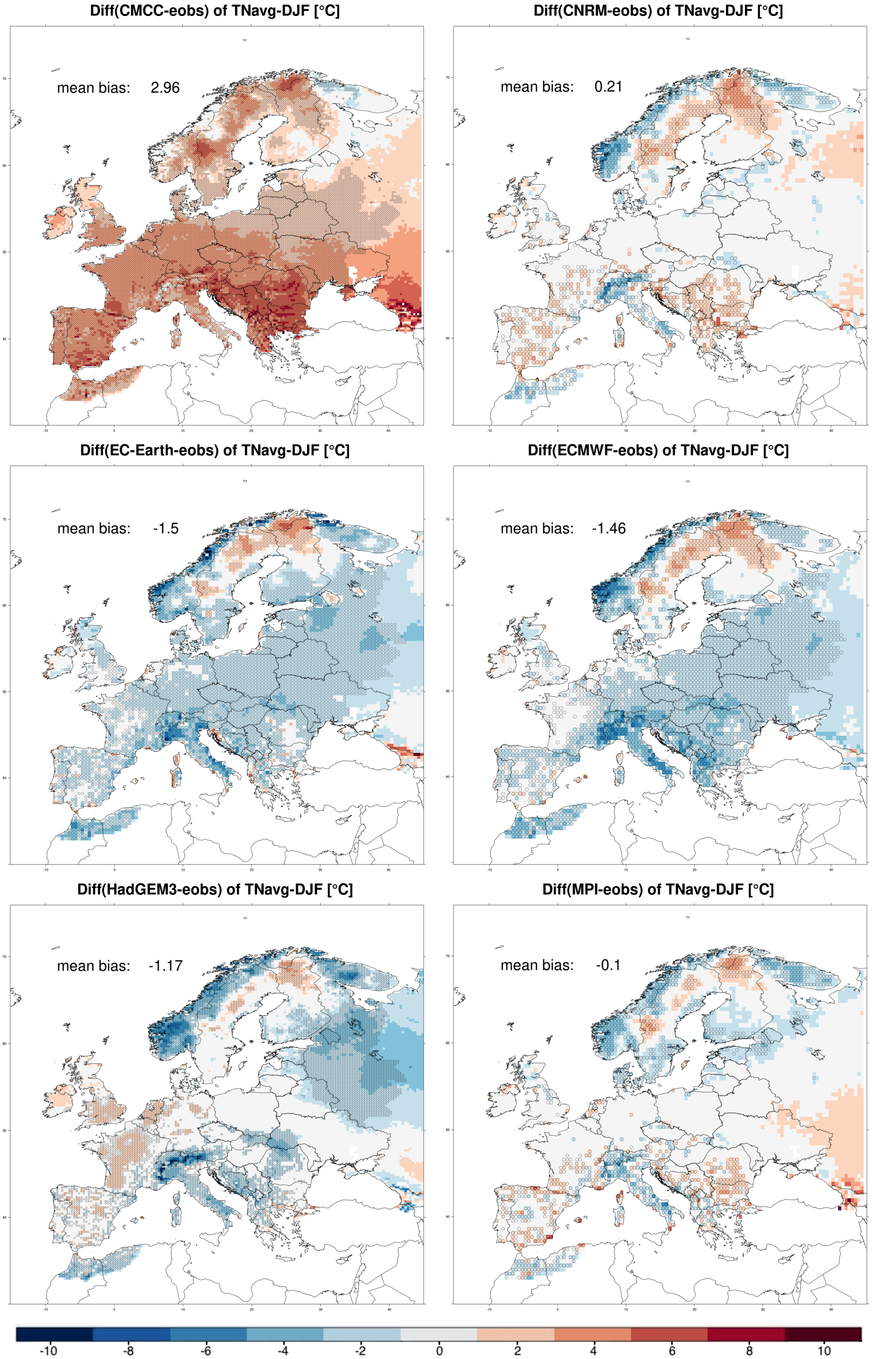
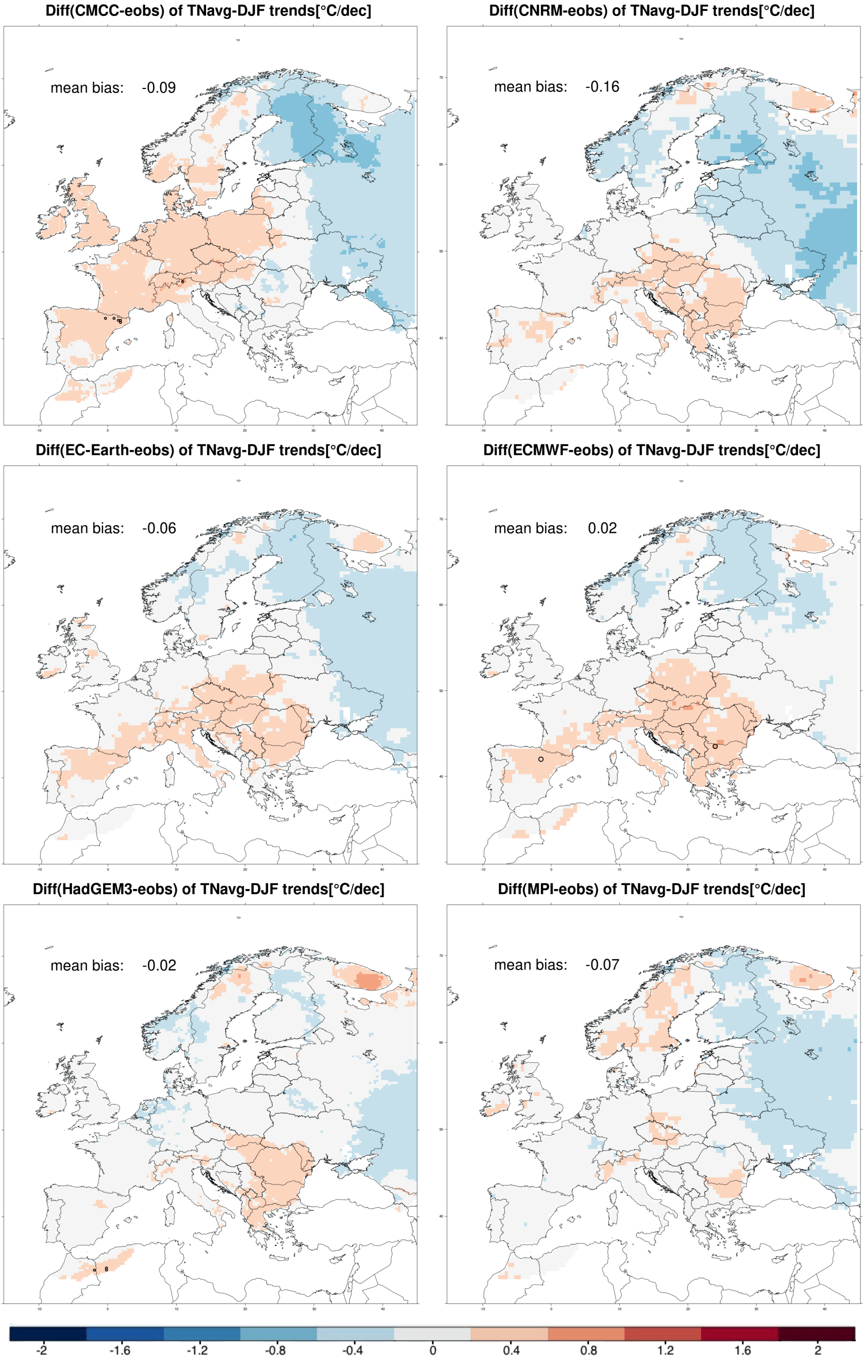

Diff(CMCC-eobs) of TNavg-DJF [°C]
mean bias: 2.96
Diff(CNRM-eobs) of TNavg-DJF [°C]
mean bias: 0.21
Diff(EC-Earth-eobs) of TNavg-DJF [°C]
mean bias: -1.5
Diff(ECMWF-eobs) of TNavg-DJF [°C]
mean bias: -1.46
Diff(HadGEM3-eobs) of TNavg-DJF [°C]
mean bias: -1.17
Diff(MPI-eobs) of TNavg-DJF [°C]
mean bias: -0.1
-10
-8
-6
-4
-2
0
2
4
6
8
10

Figure 5.1: *Difference in the winter average of TN between the HR models and E-OBS. Red indicates overestimation, blue indicates underestimation. Significant differences are indicated by small thin circles for each grid-point, which result in shaded areas.*


Diff(CMCC-eobs) of TNavg-DJF trends[°C/dec]

Diff(CNRM-eobs) of TNavg-DJF trends[°C/dec]

Diff(EC-Earth-eobs) of TNavg-DJF trends[°C/dec]

Diff(ECMWF-eobs) of TNavg-DJF trends[°C/dec]

Diff(HadGEM3-eobs) of TNavg-DJF trends[°C/dec]

Diff(MPI-eobs) of TNavg-DJF trends[°C/dec]

Figure 5.2: *Difference in the trends of TNavg-DJF for HR models. Red indicates overestimation (warmer simulated trends), blue indicates underestimation (colder simulated trends). Significant differences are indicated by black circles.*

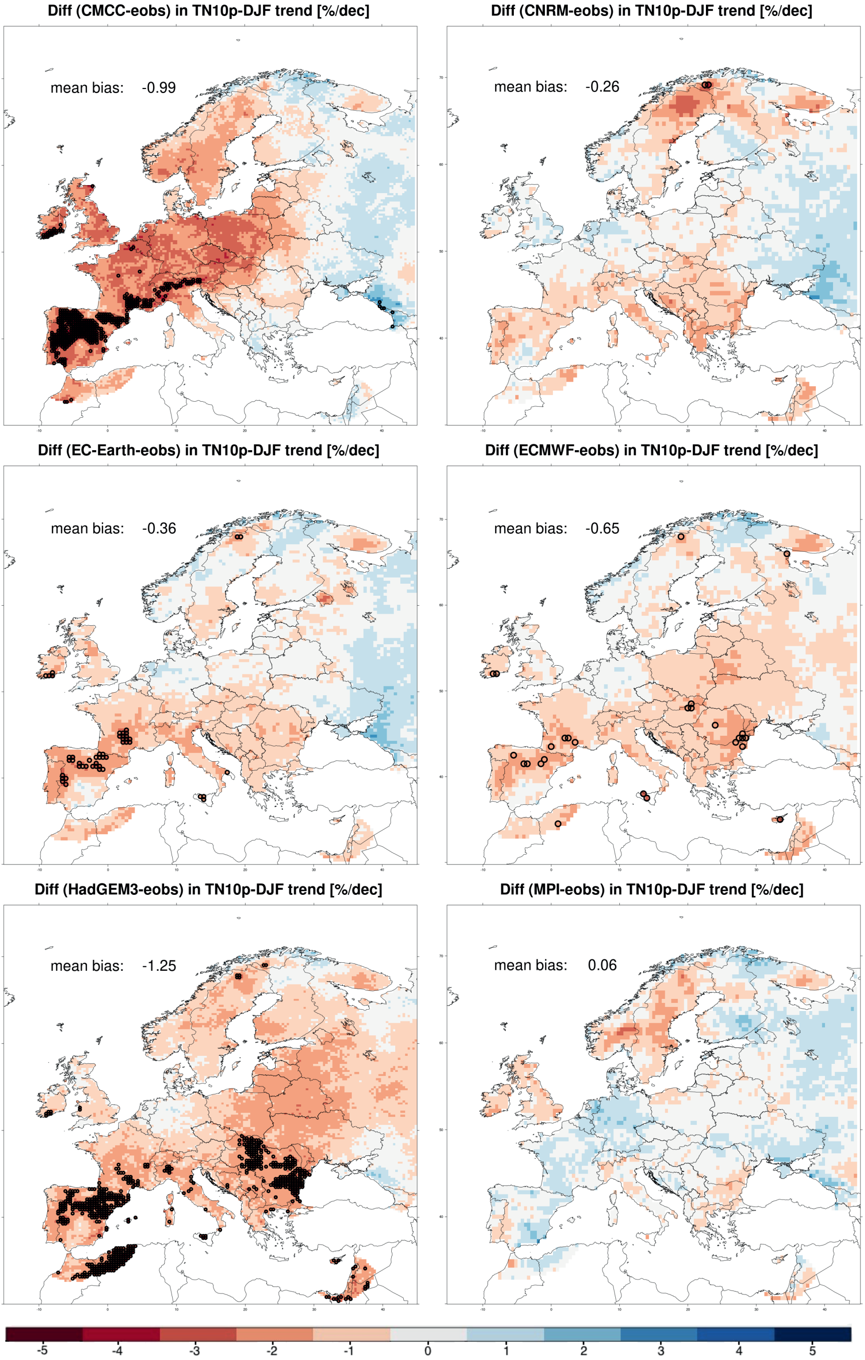
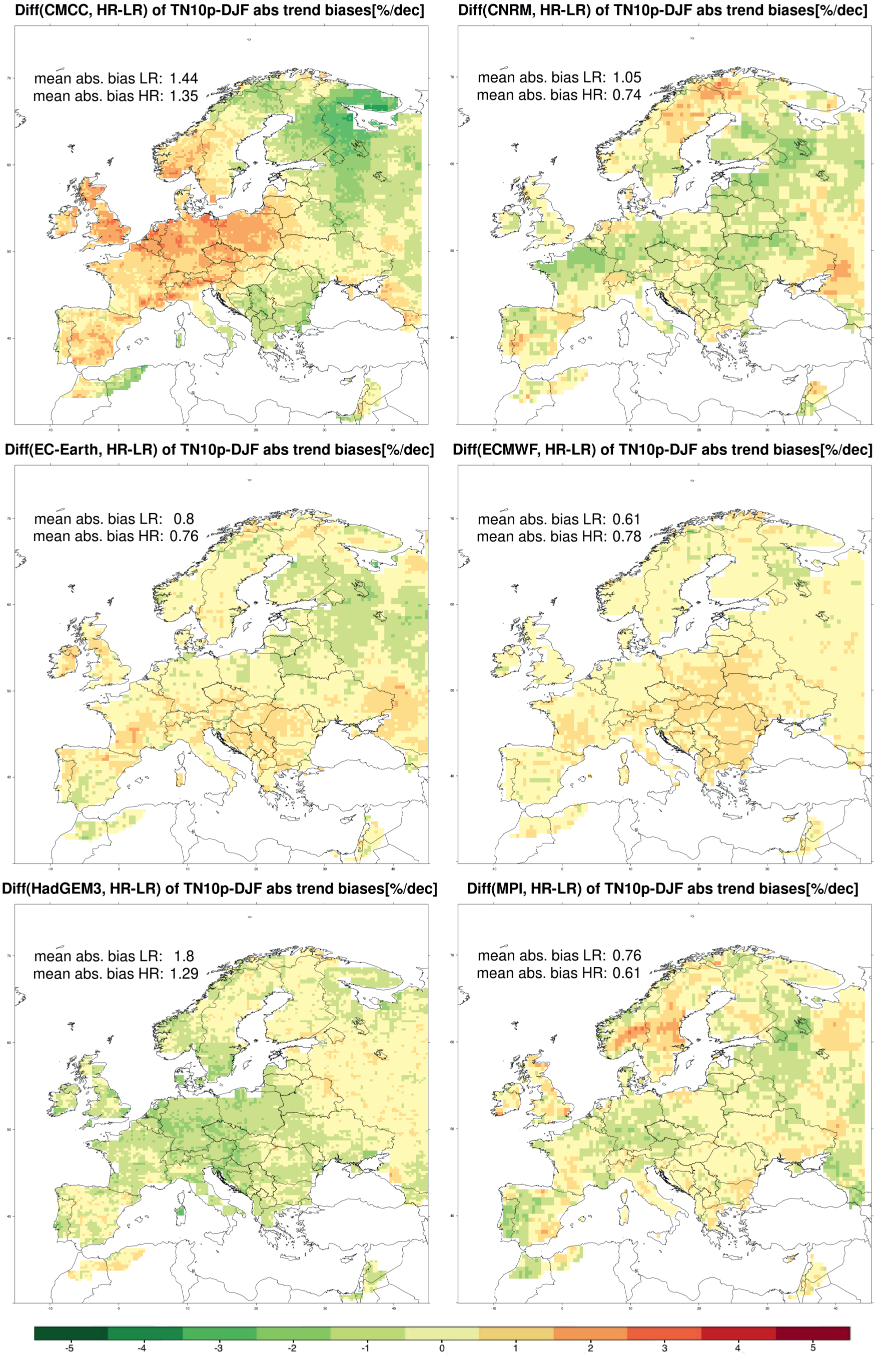

Diff (CMCC-eobs) in TN10p-DJF trend [%/dec]
mean bias: -0.99
Diff (CNRM-eobs) in TN10p-DJF trend [%/dec]
mean bias: -0.26
Diff (EC-Earth-eobs) in TN10p-DJF trend [%/dec]
mean bias: -0.36
Diff (ECMWF-eobs) in TN10p-DJF trend [%/dec]
mean bias: -0.65
Diff (HadGEM3-eobs) in TN10p-DJF trend [%/dec]
mean bias: -1.25
Diff (MPI-eobs) in TN10p-DJF trend [%/dec]
mean bias: 0.06
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 5.3: *Difference in trends of TN10p-DJF between the HR models and E-OBS. Red(blue) indicates an underestimation (overestimation) of the trend, related to a warmer (colder) trend in the model.*



Diff(CMCC, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Diff(CNRM, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Diff(EC-Earth, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Diff(ECMWF, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Diff(HadGEM3, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Diff(MPI, HR-LR) of TN10p-DJF abs trend biases[%/dec]

Figure 5.4: *Difference in absolute trend bias of TN10p-DJF between HR and LR models. Red (green) pixels indicate an increase (decrease) of the absolute trend bias, thus a better (worse) performance.*

5.3.2 Maximum temperatures

Bias in summer averages

The comparison of summer maximum temperatures, has started from the evaluation of the bias of the models. Four of them give a mean bias that has a lower absolute value compared to what is observed for winter TN, see figure 5.5. CMCC presents a large underestimation (-3.83 °C), similar, but with opposite sign, to the corresponding result for TN. The remaining 5 models show a common north-south gradient, with a warm bias along the Mediterranean and Black Sea coast and a general underestimation over Northern Europe, with different patterns and small exceptions. A large overestimation common to all models is found in Northern African regions. Nevertheless in these areas the large biases (above +10 °C) can be in part related to the high uncertainty of E-OBS, due to a lower station density.

Trends in winter averages

The difference in trends of TXavg-JJA ranges between -0.17 °C/dec (ECMWF) and +0.03 °C/dec (CMCC), see Figure 5.6. The models tend to slightly underestimate the warming of summer temperatures. This is more evident over the Mediterranean and Eastern Europe, especially for models as EC-Earthand ECMWF. These present large areas (over Italy and Balkans) where the differences are significant, implying an inaccurate reproduction of the changes in the climate of these areas. On the other side almost all the models tend to overestimate the trends over Southern Scandinavia, especially over Norway.

Trends on warm extremes

Figure 5.7 helps evaluating the reproduction of TX90p-JJA, which describes warm temperature extremes. The results show a large underestimation of the trends for EC-Earth(-0.73 %/dec), ECMWF (-0.59 %/dec) and HadGEM3 (-0.56 %/dec). In all cases stronger trends, consistent with what found for the trends on the averages, are simulated over Norway and Sweden. The overall bias in the MPI model is very small, but the underestimation of trends in south-eastern Europe apparently compensates the overestimation of trends in north-western Europe. This aspect (as found for the simple seasonal averages as well) is simulated by most of the models, indicating a general tendency to reproduce lower trends of warm extremes on the Mediterranean and Black Sea region and slightly larger ones around Northern Sea. In these areas large significant differences are found in particular for EC-Earth3, ECMWF and HadGEM3.

Figure 5.8, showing the difference in absolute trend biases between the HR and LR model configuration, does not show a common pattern. Best improvement in the passage from

HR to LR is for MPI (-0.16 %/dec), while HadGEM3 presents the largest worsening (+ 0.25%/dec). No particular geographical structures are found in this case. This result indicates that the reproduction of trends of warm extremes with High Resolution models hasn't considerably improved over Europe for most of the models.

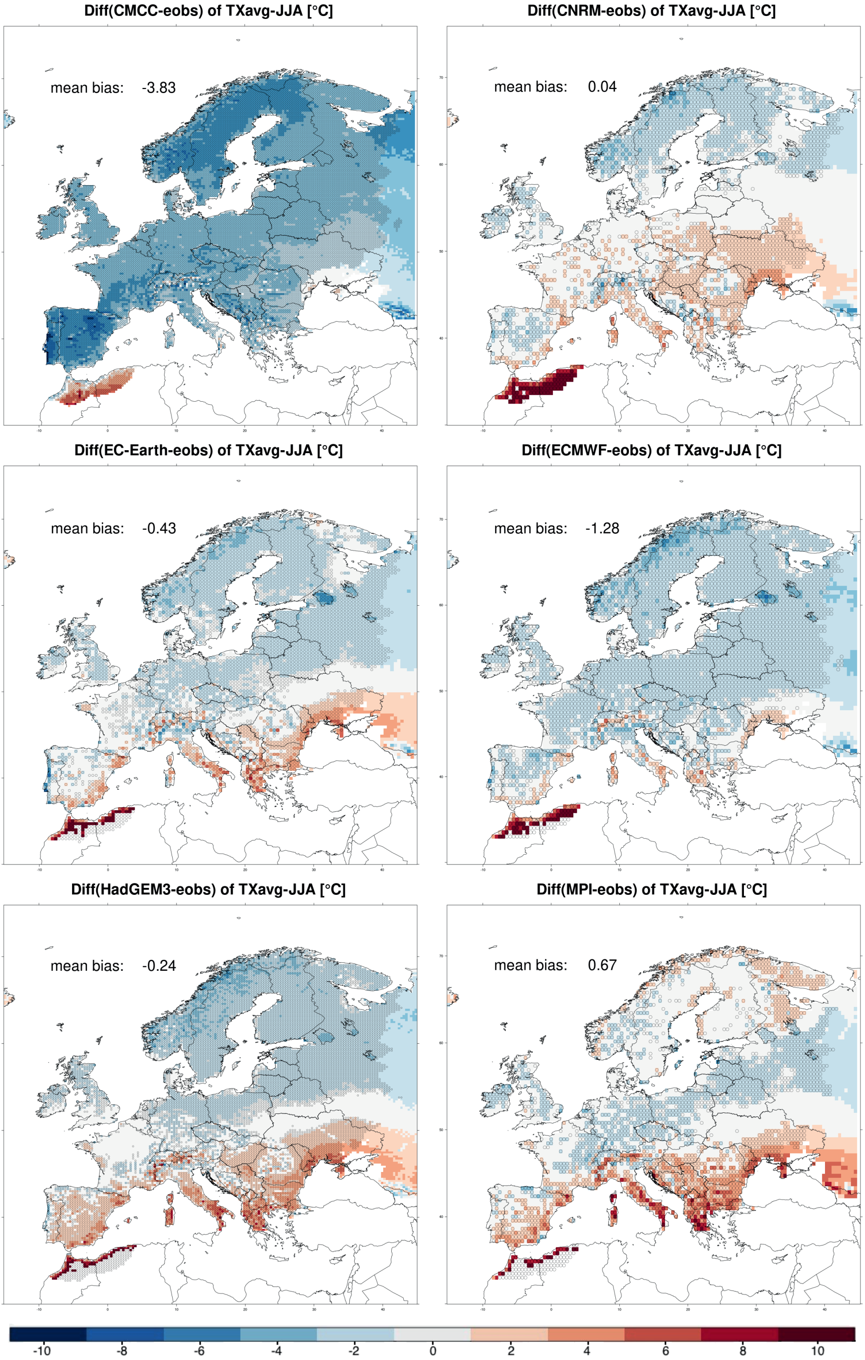

Diff(CMCC-eobs) of TXavg-JJA [°C]
mean bias: -3.83
Diff(CNRM-eobs) of TXavg-JJA [°C]
mean bias: 0.04
Diff(EC-Earth-eobs) of TXavg-JJA [°C]
mean bias: -0.43
Diff(ECMWF-eobs) of TXavg-JJA [°C]
mean bias: -1.28
Diff(HadGEM3-eobs) of TXavg-JJA [°C]
mean bias: -0.24
Diff(MPI-eobs) of TXavg-JJA [°C]
mean bias: 0.67
-10
-8
-6
-4
-2
0
2
4
6
8
10

Figure 5.5: *Same as figure 5.1 but for TX-JJA.*

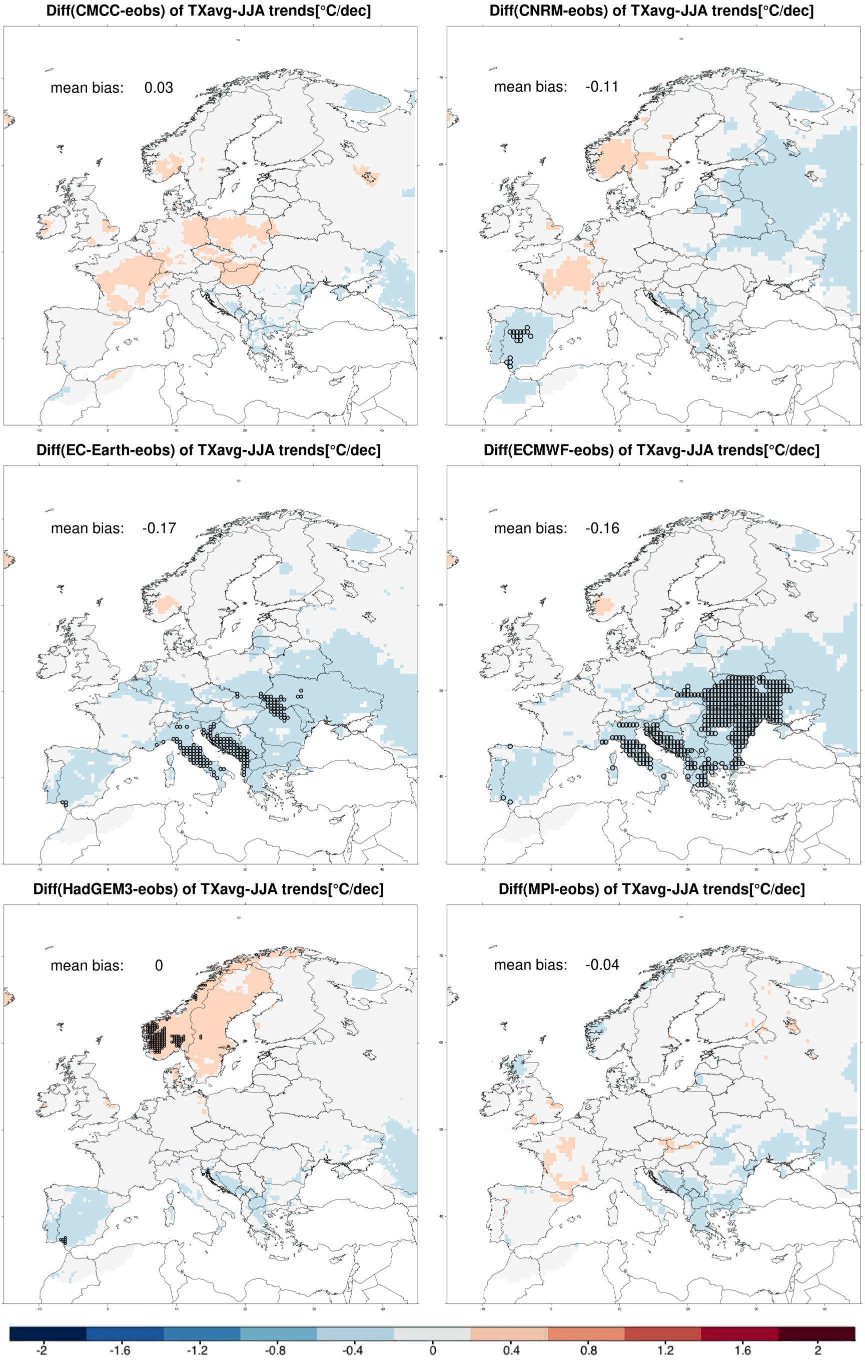

Diff(CMCC-eobs) of TXavg-JJA trends[°C/dec]
mean bias: 0.03
Diff(CNRM-eobs) of TXavg-JJA trends[°C/dec]
mean bias: -0.11
Diff(EC-Earth-eobs) of TXavg-JJA trends[°C/dec]
mean bias: -0.17
Diff(ECMWF-eobs) of TXavg-JJA trends[°C/dec]
mean bias: -0.16
Diff(HadGEM3-eobs) of TXavg-JJA trends[°C/dec]
mean bias: 0
Diff(MPI-eobs) of TXavg-JJA trends[°C/dec]
mean bias: -0.04
-2
-1.6
-1.2
-0.8
-0.4
0
0.4
0.8
1.2
1.6
2

Figure 5.6: *Same as figure 5.2 but for TX-JJA.*

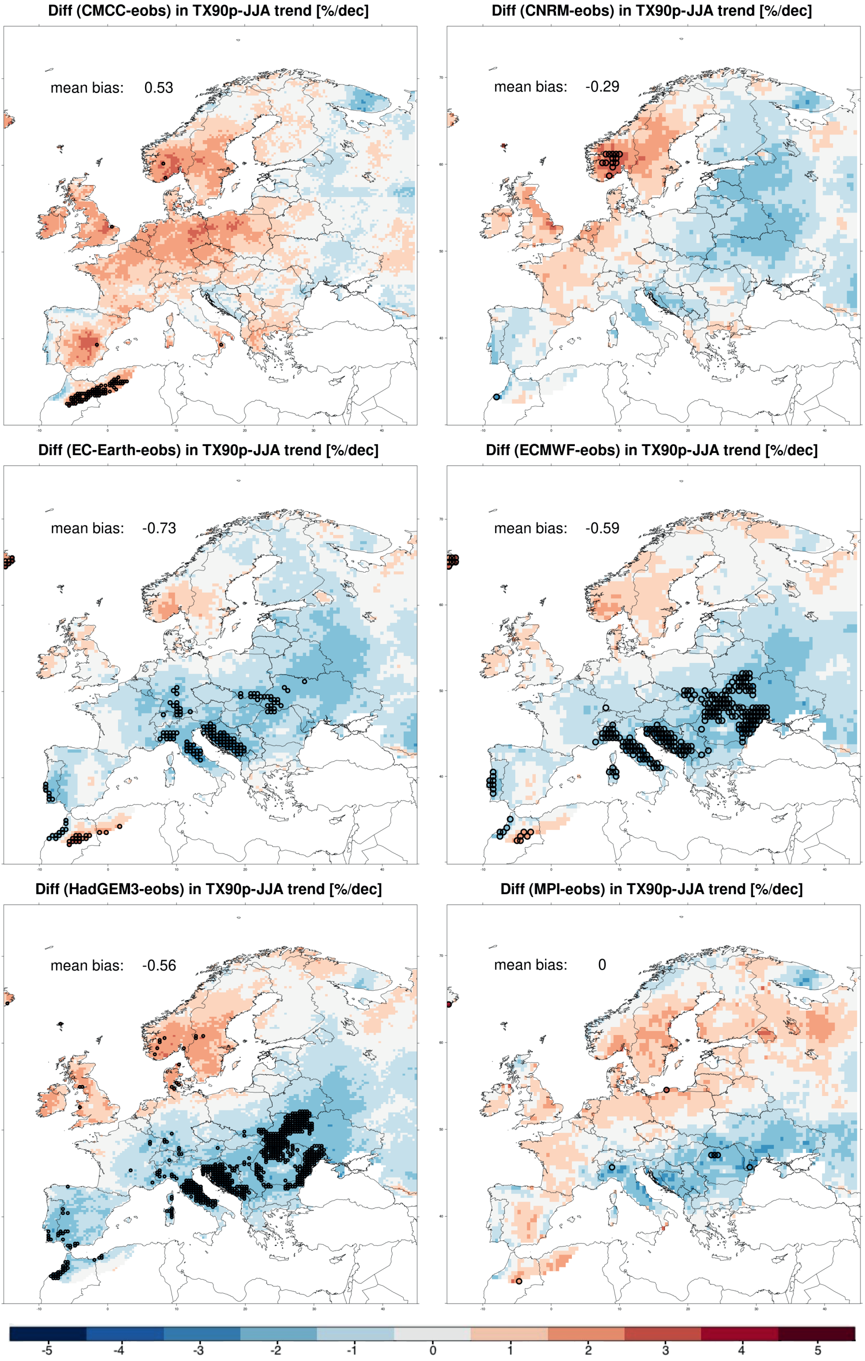
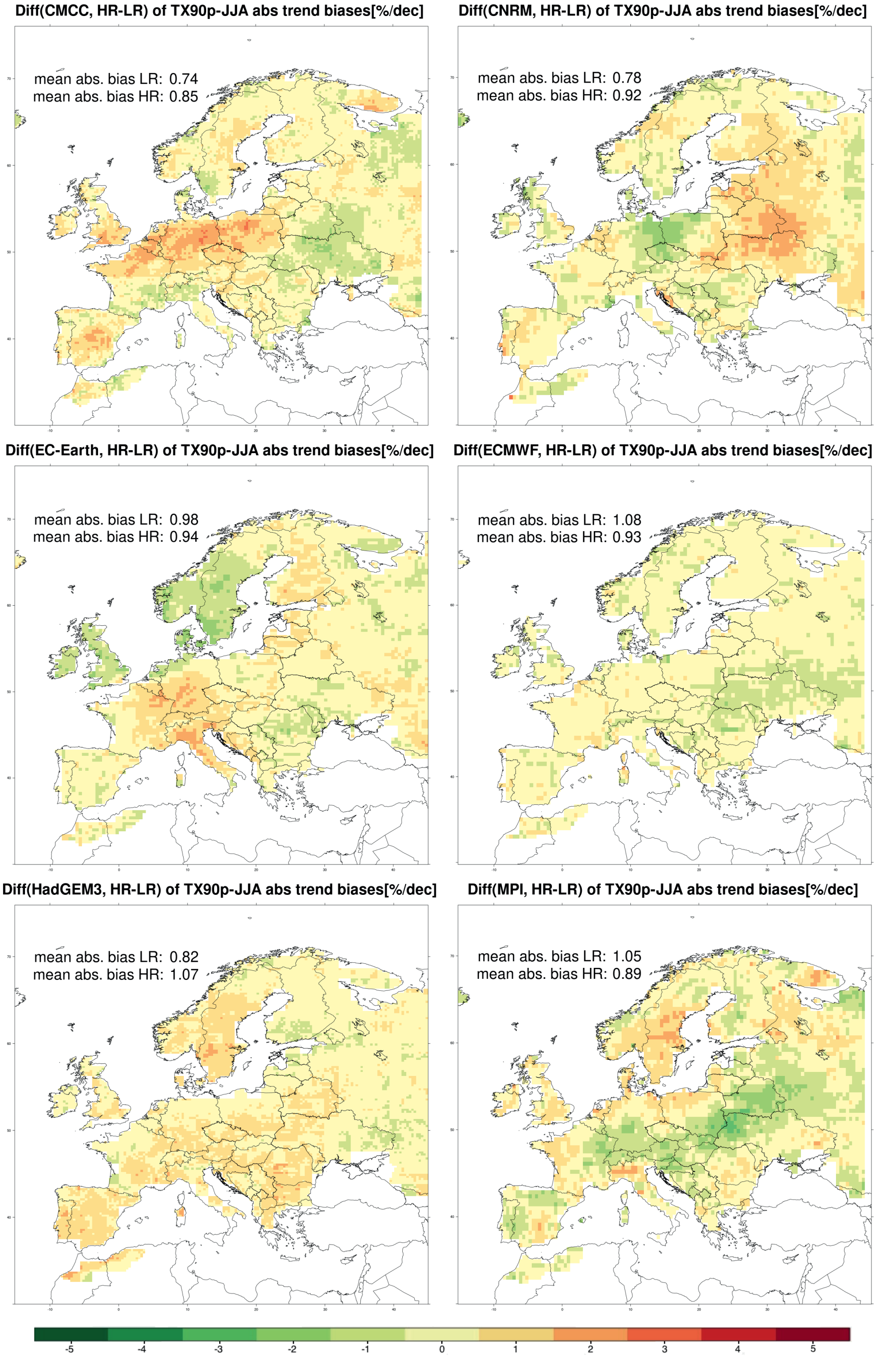

Diff (CMCC-eobs) in TX90p-JJA trend [%/dec]
mean bias: 0.53
Diff (CNRM-eobs) in TX90p-JJA trend [%/dec]
mean bias: -0.29
Diff (EC-Earth-eobs) in TX90p-JJA trend [%/dec]
mean bias: -0.73
Diff (ECMWF-eobs) in TX90p-JJA trend [%/dec]
mean bias: -0.59
Diff (HadGEM3-eobs) in TX90p-JJA trend [%/dec]
mean bias: -0.56
Diff (MPI-eobs) in TX90p-JJA trend [%/dec]
mean bias: 0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 5.7: *Difference in trends of TX90p-DJF between the considered models and E-OBS. Red(blue) indicates an overestimation (underestimation) of the trend, related to a warmer (colder) trend.*



Diff(CMCC, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Diff(CNRM, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Diff(EC-Earth, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Diff(ECMWF, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Diff(HadGEM3, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Diff(MPI, HR-LR) of TX90p-JJA abs trend biases[%/dec]

Figure 5.8: *Same as figure 5.4 but for TX90p-JJA.*

5.4 Summary and Conclusions

Six models in their High (HR) and Low Resolution (LR) versions have been compared (over the 1970-2014 period) to E-OBS.hom, a version of the gridded dataset E-OBS based on homogenized daily series (each covering at least 1970-2014) of observed temperatures. The analysis has been performed focusing first on the biases and the trend biases of mean values of winter minimum temperatures (TNavg-DJF) and mean values of summer maximum tempratures (TXavg-JJA) and then on two ETCCDI [ETCCDI, 2009] defined indices. These are the number of days with minimum temperatures below the 10th percentile of winter values ('cold nights', TN10p-DJF) and the number of days with maximum temperatures exceeding the 90th percentile of summer values ('warm day-times', TX90p-JJA). The percentile thresholds have been calculated using the 1981-2010 period. After the calculation of the trends of the considered indices, for those models with more than one ensemble member (see table 5.1), the ensemble mean has been calculated. For each grid-point, average values and trends in the models have been compared to observations and an assessment is made of the difference between the HR and LR model versions.

For both winter-mean TN as summer-mean TX strong biases have been found in the simulations, with the strongest ones for CMCC. This model shows mean bias of +2.96°C/dec for TNavg-DJF and -3.83°C/dec for TXavg-JJA, indicating an underestimation of the seasonal cycle all over the continent. The other models present smaller biases (averaged over Europe). Nevertheless common patterns are found, such as an underestimation of winter minimum temperatures over Italy and Norway and a shared overestimation in the north of Sweden and Finland. This last issue may be related to a lack of snow coverage simulated by the models, as suggested by [van Oldenborgh et al., 2009] and [Diffenbaugh et al., 2013].

As for maximum temperatures in summer: the models share a common North-South gradient in the bias, with warmer values along the European coasts of the Mediterranean. This may be related to excessive moisture in Northern Europe and a lack of moisture in the Southern sector [Seneviratne et al., 2006; van Oldenborgh et al., 2009; Lorenz et al., 2010]. Evaluation of results for TXavg-JJA shows a slight overestimation of trends for HR compare to E-OBS on Northern Europe and an underestimation on Southern Europe, especially over Italy and the Balkans. This pattern is consistent with what found by [Bhend and Whetton, 2013], which detected an underestimation of trends of average summer TX over Europe in CMIP5.

In Southern Europe, the combination of an excessively large negative bias in summer maximum temperatures with a too weak increase in the seasonal average and with a much weaker (compared to observations) increase in the extreme indices points to issues in the representation of soil moisture in the models. In a climate which is too warm the soil can be expected to lack more moisture than in cooler conditions due to enhanced evaporation. Once the soil is dry the radiation balance is shifted to a state where sensible heat is dominant over latent heat. Under boundary conditions where the incoming energy flux (due to increase of green house gases) raises, this implies a further increase in sensible heat and surface warming. Nevertheless in conditions of moist soil, the simulated warming trend in temperatures would be even stronger, due to the shift from latent, thus getting close to the observed conditions [Seneviratne et al., 2006; van Oldenborgh et al., 2009; Lorenz et al., 2010; Min et al., 2013].

The most interesting aspects on the trends in winter temperatures, figure 5.2, is the simulation of colder trends in Eastern Europe (excluding the Kola peninsula) common to all models. Such anomaly might be linked to too small simulation of the reduction trend of snow coverage compared to the observations [van Oldenborgh et al., 2009] and suggests to conduct an inspection on the performances of models on this particular variable, which will be subject of future studies.

The too warm simulated trend on the peninsula of Kola is found in the trends on TNavg-DJF and TN10p-DJF (as an underestimation of the number of days below the 10th percentile) and is related to E-OBS station density issues. The only series with observed values in the area (Krasnoshelye) starting before 1970 has missing data between 1972 and 1980. The interpolation of data coming from series in surrounding stations, in the case of TN, brings to higher values in the 1972-1980 compared to the following years, introducing a too cold trend that doesn't take place in the models. This behavior, limited to only one series, motivates the ECA&D group to work on further data collection and in increasing the station density in this and other areas. This will allow to increase the quality of the interpolation and avoid such criticisms.

Trends in extreme values have presented several anomalies, often with common geographical patterns among the models. While the underestimation of trends of TNavg-DJF simulated over Eastern Europe is found also for TN10p-DJF five models indicate an underestimation of the percentage of cold days, thus warmer trends, over Southern Europe. At the same time the underestimation of percentage of warm day is found for the trends of TX90p-JJA, indicating colder trends than the observed ones, consistent with the findings of [Min et al., 2013] for CMIP5.

The combination of these two aspects indicates that around the Mediterranean the model trends in the tails of the distribution are closer to each other than what is observed. Therefore in Southern Europe the distribution of simulated daily temperatures tends to get narrower compared to the distribution of observed daily temperatures, underestimating the intensity of the extremes, especially the warm ones.

As a last step, the analysis of the absolute trend bias evolution in the models from LR to HR does not show a general improvement. Each model presents different patterns and diverse behavior in terms of change of mean absolute trend bias. Nevertheless this index decreases for TN10p-DJF in almost all models (but ECMWF), indicating a better improvement compared to what id found for TX90p-JJA, where only 3 methods slightly improve and the other ones present worsening up to +0.25%/dec.

Finally, it appears that the new high resolution models, even though they do not significantly increase or decrease their absolute bias on the trends of the extremes, still present some criticisms especially on the area of the Mediterranean. In this region the most serious discrepancy to observations is the large underestimation of the increasing trends of warm extremes. Considering the high economic and societal vulnerability of these areas to very warm events in summer and the importance of the prediction of heatwaves intensity and frequency for the next decades, it is fundamental to improve the simulation of these phenomena and of their projections to future decades.

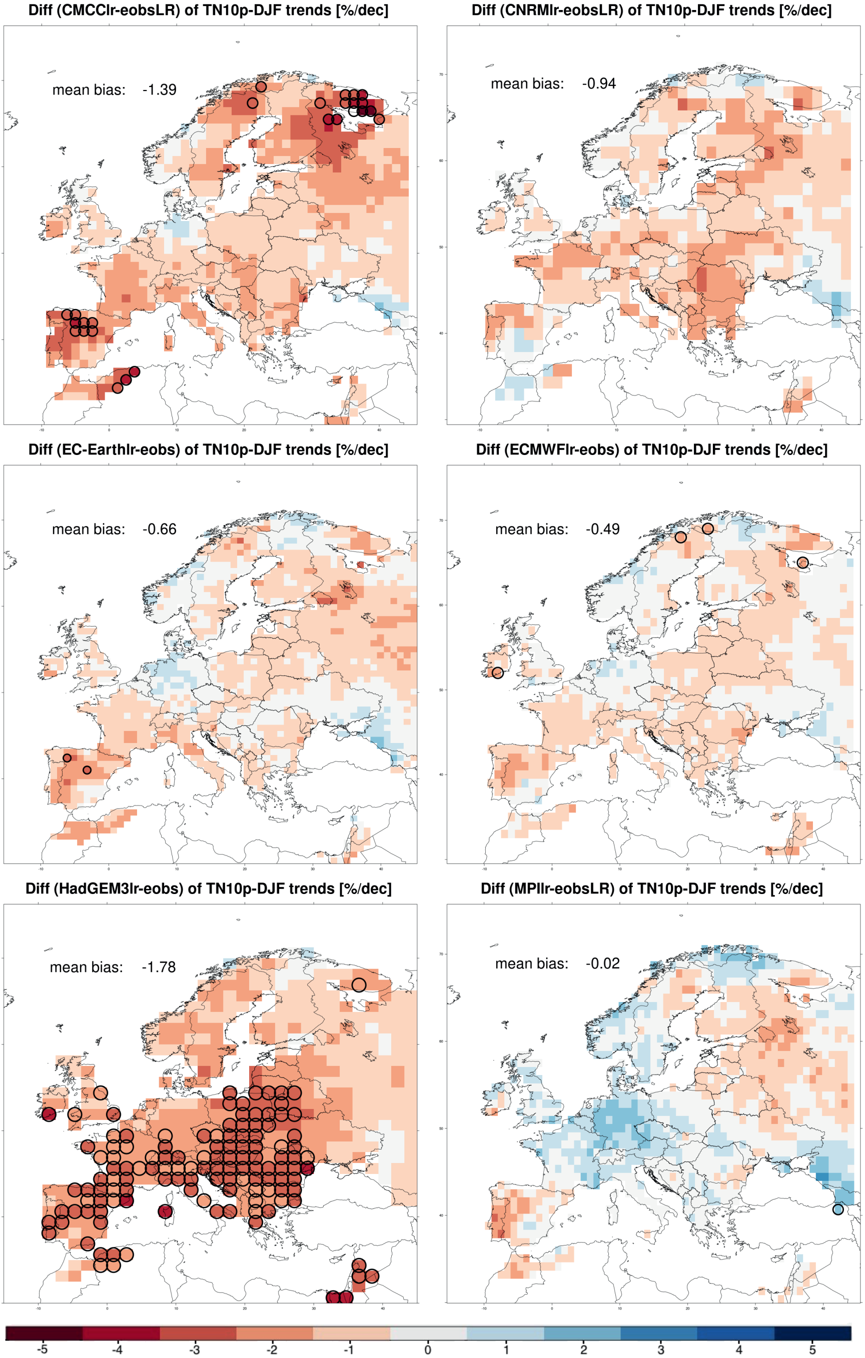
Acknowledgements

We acknowledge funding from the PRIMAVERA project, funded by the European Union's Horizon 2020 programme under Grant Agreement no.641727. The E-OBS observational dataset is made available through the European Climate Assessment & Dataset (www.ecad.eu).

5.A Performance of the models in their Low Resolution version

The LR version of the models tend to underestimate the trend in TN10p-DJF. Figure 5.9 displays that all the mean biases are negative. This indicates a general overestimation of the warming trend, with a few exceptions for some models in certain areas, e.g. MPI in Central Europe. Only for a small part of the grid-points the difference is significant, i.e. the 95% confidence level ranges of the models and E-OBS trends don't superimpose.

Performance of LR models on trends of TX90p-JJA present different patterns. In general trends on Eastern Europe are underestimated (colder trends than observed). Nevertheless the mean biases indicate that, contrarily to what is seen for the other models, CMCC and CNRM slightly overestimate the trends.



Diff (CMCClr-eobsLR) of TN10p-DJF trends [%/dec]

Diff (CNRMlr-eobsLR) of TN10p-DJF trends [%/dec]

Diff (EC-Earthlr-eobs) of TN10p-DJF trends [%/dec]

Diff (ECMWFlr-eobs) of TN10p-DJF trends [%/dec]

Diff (HadGEM3lr-eobs) of TN10p-DJF trends [%/dec]

Diff (MPIlr-eobsLR) of TN10p-DJF trends [%/dec]

Figure 5.9: *Difference in the trends of TX10p-DJF between the LR models and E-OBS. Red (blue) indicates that the models have more negative (positive) trends in the number of days above the 10th percentile, related to warming (cooling) trends. Circles indicate significant differences.*

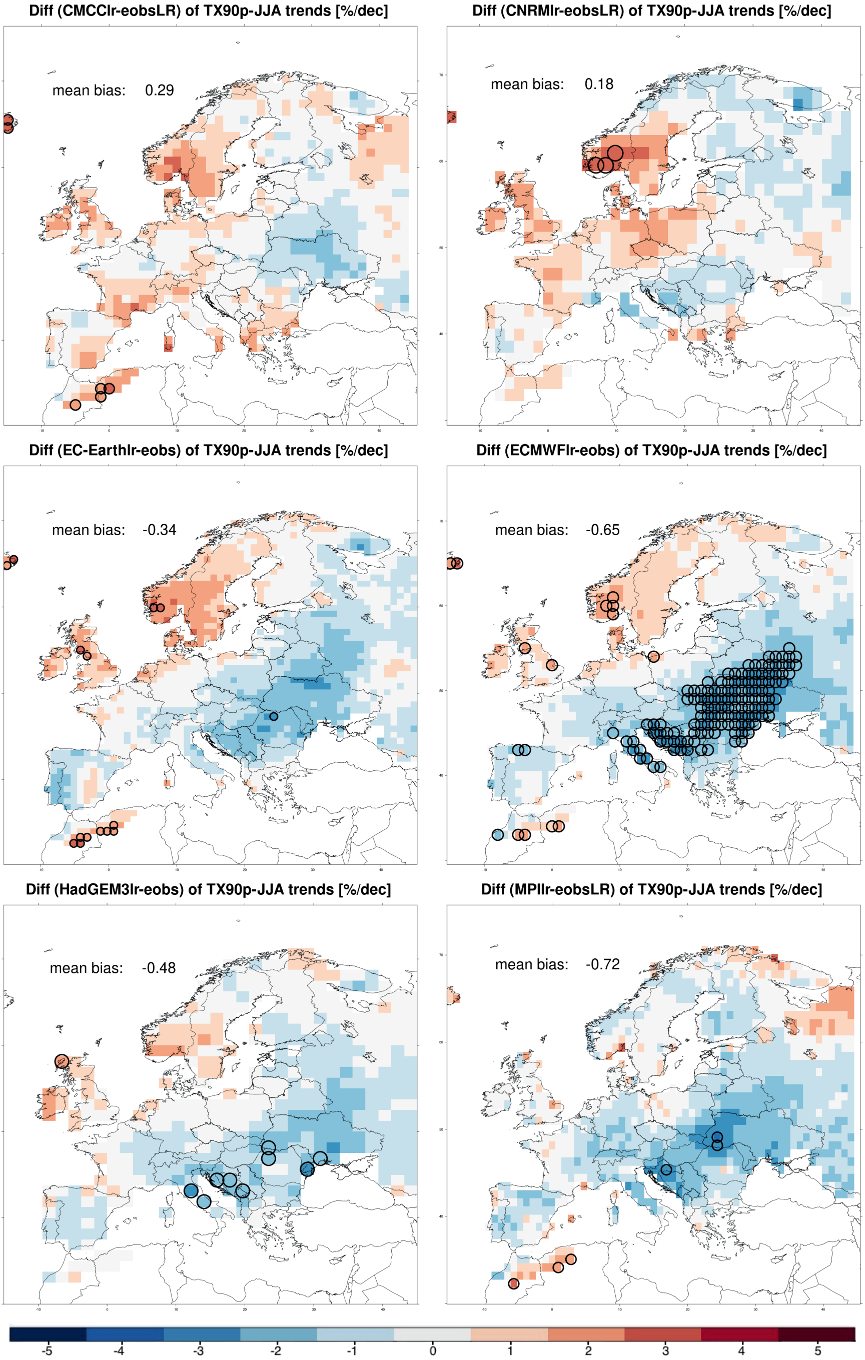

Diff (CMCClr-eobsLR) of TX90p-JJA trends [%/dec]
mean bias: 0.29
Diff (CNRMlr-eobsLR) of TX90p-JJA trends [%/dec]
mean bias: 0.18
Diff (EC-Earthlr-eobs) of TX90p-JJA trends [%/dec]
mean bias: -0.34
Diff (ECMWFlr-eobs) of TX90p-JJA trends [%/dec]
mean bias: -0.65
Diff (HadGEM3lr-eobs) of TX90p-JJA trends [%/dec]
mean bias: -0.48
Diff (MPIlr-eobsLR) of TX90p-JJA trends [%/dec]
mean bias: -0.72
-5
-4
-3
-2
-1
0
1
2
3
4
5

Figure 5.10: *Difference in the trends of TX90p between the LR models and E-OBS. Red (blue) indicates that the models have more positive (negative) trends in the number of days above the 90th percentile, related to warming (cooling) trends. Circles indicate significant differences.*

Chapter 6

Synthesis

6.1 Aims of the research

The present thesis contributes to the ability of the scientific community to improve the monitoring of changes in the extreme climate events, based on in-situ observations. These measurements, compared to other systems such as satellite measurements or reanalyses, present longer temporal coverage (up to more than 200 years). Furthermore they are direct result of an observation of real climatic conditions and are not the result of inversion or modelling processes.

The use of daily temperature series is compromised by human interventions to the stations, such as changes of position within a site, changes in the instrumentation or relocations of the site itself. These introduce biases, which can be sudden or gradual, often have similar amplitudes as the climatic signal [Aguilar et al., 2003; Caussinus and Mestre, 2004; Begert et al., 2005; Della-Marta and Wanner, 2006; Menne and Williams Jr, 2009]. The presence of such inhomogeneities involves a large portion of raw observations [Tuomenvirta, 2001; Hartmann et al., 2013; Li et al., 2016]. Therefore thorough homogeneity assessments and adjustments are mandatory before any climatological analysis. This will convey further robustness to the scientific evidences of climate change [Domonkos, 2011; Trewin, 2013; Lindau and Venema, 2016]. In addition, the homogeneity of station series is fundamental for the reliability of derived products as gridded datasets [Venema et al., 2013; Cornes et al., 2018].

In order to provide a reliable source for climate change assessment in Europe, this project has aimed at developing a homogenization method to be applied to the station data in the European Climate Assessment and Dataset (ECA&D). The ECA&D collects thousands of temperature series from all over Europe (Figure 3.1), making it one of the largest databases of meteorological in-situ observations on the continent. This dataset is widely used within the research community, as demonstrated by the more than 1400 citations obtained by Klein Tank et al. [2002] (status January 2020), which first introduced and described the dataset. Furthermore, the station records of ECA&D are used in the construction of the E-OBS gridded dataset, which has been of great interest for the scientific community (more than 2200 citations for Haylock et al. [2008] and more than 55 for the recently introduced Cornes et al. [2018]).

The history of the measurements in ECA&D (*metadata*), like changes in location or instrumentation, is often poorly documented. In addition, the sheer size of the dataset simply makes it excessively time consuming to involve careful analyses of metadata. Thus, the need is to develop a completely automated procedures, able to autonomously track the inhomogeneities (break detection) and adjust them (adjustment calculation) [Dienst et al., 2017; Delvaux et al., 2018].

On one hand the break detection has to be able to identify the change points with the lowest possible rate of false positives. On the other side, the adjustment calculation needs to determine which factors to add to the values in order to remove non-climatic signals from the series. These routines need to work efficiently on the ECA&D. Hence, the procedure has to be flexible enough to manage wide spectrum of inhomogeneity signals and areas with very sparse station density. A further aim of the study has been to implement the procedure with an approach that manages the data without employing parametrizations (e.g. fit to models, linear and non-linear regression), so that the adjustments derive from real observed signals and are not filtered with pre-defined analytical functions.

In several cases the series (often located in urban areas) are ended and new ones are simultaneously started in a nearby location (usually in the airport or in a rural area). These series are gathered in order to generate long series, ideal for thorough historical assessments. Long records are constructed through a process called 'blending', in which series of neighbouring stations are combined into longer series. Since this process generates further (and more complex) inhomogeneities, the adjustment calculation procedure applied to the undocumented inhomogeneities needs to be adapted to the blended series.

A corner stone of this thesis has been the focus on extreme temperatures, due to their higher impact on society and economy compared to average values [Donadelli et al., 2017]. A thorough trend analysis on station data has been among the aims of this dissertation, including assessments on the effects brought by the homogenization process and inspections of the possible phenomena that drive the observed changes in the climate.

Finally, it is fundamental that any homogenization method is validated in order to assess the quality of its products and, once this is done, that powerful examples of the possible uses of the new homogenized data are provided. By demonstrating the usefulness and solidity of the developed routines and the obtained products, the aim is to increase trust and confidence in the user community.

6.2 Results and Discussion

6.2.1 The new homogenization method

In the frame of this thesis a homogenization procedure has been developed following the requirements of the goals listed above: being fully automated, conservative (avoiding false positives and overcorrections), heuristic (favouring the real signal against models and regressions) and flexible (able to handle all realistic inhomogeneities and all station density conditions). The break detection, based on the work of Kuglitsch et al. [2012], checks the series with three different methods [Caussinus and Mestre, 2004; Wang et al., 2007; Toreti et al., 2012] and looks for agreement of at least two of them. In the three cases reported in Chapter 2, the metatadata show good agreement with the detected breaks.

Such findings indicate that break detection presents a good accuracy, even though some breaks don't correspond to a documented change in location or equipment. Unfortunately it's not possible to determine if these effectively are false positives. Although these are impossible to avoid, the fact that two break detection methods out of three indicate them as change points reassures on the goodness of the result.

Each detected break divides the target series into two sub-series whose data, month by month, are binned into quantiles and compared, identifying the difference in the shape of the distributions before and after the break. This difference is generated by climatic and artificial signals. Since the first needs to be preserved in the final homogenized series, the second one has to be isolated. For this reason the climatic signal is identified by considering homogeneous reference series, whose quantile differences are subtracted from the one related to the target. This approach, whose details are explained in Chapter 2, allows to elaborate customized adjustments for extreme values. In addition it is flexible enough to handle situations with a low station density while remaining computationally efficient in case of a very high station density, as proved in its application on different benchmark datasets in Chapter 4.

A strong point of the quantile matching procedure described in this work is its heuristic approach. The procedure developed in this project takes the natural signal as it comes from the measurements, working on the real observed distribution of the values This contrasts with other methods that determine the values of the adjustments with parametrizations of the temporal series [Menne and Williams Jr, 2009] or regression models applied to the probability density function [Della-Marta and Wanner, 2006; Mestre et al., 2011]. Such aspect conveys flexibility, allowing that any (realistic) shape of the distribution is handled and adequately managed, without forcing the data into expected behaviours or parametrized patterns. These benefits involve especially the extreme values, whose adjustment factors are calculated without any influences from the first moment of the distribution. Furthermore in the whole process the necessity not to over-adjust the data has always been taken into consideration. This conservative approach has, for example, been implemented with the smoothing of the adjustments and with the check of negative slopes in the adjustments (see Appendix A of Chapter 2), which have respectively aimed at removing noise-induced signals and at not varying, during the process, the rank of a value within its sample.

Similarly to previous studies [Brunet et al., 2006; Kuglitsch et al., 2009; Syrakova and Stefanova, 2009; Nemec et al., 2013; Mamara et al., 2014] only 10% of the series of ECA&D were already homogeneous. As expected from the large presence of relocations from cities to rural areas [Brunet et al., 2006; Nemec et al., 2013; Yosef et al., 2018] a slight majority of the adjustments applied to the earlier parts of the series (in order to be consistent with the latest ones) are negative. This due to the removal of the Urban Heat Island Effect from the (earlier) parts related to the urban stations, which, lowering the values, causes slightly warmer trends.

By consequence cooling biases are introduced to the series, these require to be adjusted lowering the values related to the urban stations. The case studies display a clear improvement in the quality of the series (Figures 2.9,2.13,2.17).

The new homogenized version of the original series in the ECA&D is the input for the updated blending procedure, enriched with a new module for the removal of the duplicates generated by the blending itself. (Chapter 3). The blended series are then homogenized with a special version of quantile matching able to manage individual isolated values and longer portions of data. The case studies explained in this Chapter demonstrate again the strength of the quantile matching homogenization, highlighting the different results between average and extreme values and presenting the capability to handle cases with complex blending.

The combination of these two steps generates a data set with long and homogeneous series, these are an extremely precious tool for the inspection of the long-term climate of Europe from the historical and geographical points of view. One of the aims of this thesis has been to identify the contributions of homogenization to the general statistical characteristics of the dataset, exploring the benefits and eventual unexpected atypical behaviours. The network average trends of annual mean, 5th and 95th percentiles present a limited change (between +0.01 and +0.02 °C/dec). However, anomalous trends that could be observed in the raw dataset are now absent. The same effect can be seen in the reduction of the range of trends of the series (Figure 2.16) and in the relevant increase of geographical consistency (Figures 2.14, 3.12). This proves that, while the network averages don't show relevant changes, the benefits for single stations and for the description of the local climate are clear, as stated in several previous studies [Tuomenvirta, 2001; Brunetti et al., 2006; Li et al., 2016]. In addition no particular patterns in the change of trends of series are observed (Figures 2.5,3.12), confirming that the aim of this thesis (and homogenization procedures in general) is not to intentionally increase the trends, but to just provide data which are immune from non-climatic interferences.

The Quantile Matching has been, together with five more homogenization methods, evaluated against a benchmark dataset in Chapter 4. This has been done against two benchmark datasets, whose construction has been performed concatenating portions of homogeneous neighbouring series, This process allows to create series affected by a plausible relocation-like signal, tests, whose unperturbed version is known, benchmarks. The evaluation has been made using widely known metrics such as the Root Mean Square Error, the calculation of the percentage of adjusted data and the formulation of a new indicator that is able to calculate how much the homogenized version gets close to the benchmark characteristics. The application of Quantile Matching to a low-station-density benchmark dataset has shown that a lower percentage of non-homogenizable values (Table 4.3), a better performance in terms of root mean square error and a very satisfactory reproduction of the trends of average and extreme values (Figure 4.13). This has demonstrated that Quantile Matching performs on the level of widely known and prestigious homogenization methods as SPLIDHOM [Mestre et al., 2011] and is robust enough to handle both very favourable and very challenging conditions.

6.2.2 Observed changes in European climate

The new dataset of homogenized blended series allows to inspect trends on the long term over Europe, focusing on extreme values and on eventual changes in the width of the distribution. Strong warming trends in winter temperatures over Alps and Eastern Europe have been identified, similarly to what observed in other studies [Simolo et al., 2010; Osadchyi et al., 2018; Krauskopf and Huth, 2019]. In addition the cold tail (10$^{\text{th}}$ percentile) has shown warmer trends than the warm tail indicating a relevant narrowing (i.e. cold values have experienced a steeper increase). These two facts are linked to the reduction in snow coverage, that affects the cold part of the distribution reducing the number of days with very cold night-time temperatures. On the other side, steep trends in both the 10$^{\text{th}}$ and 90$^{\text{th}}$ percentile of summer maximum temperatures in Southern and Western Europe have been found. This is consistent to earlier findings of several works that focused on the area [Brunet et al., 2006; Della-Marta et al., 2007; Simolo et al., 2010; Delvaux et al., 2018; Fioravanti et al., 2019] that have observed how this area, and especially the Mediterranean have been involved by increased length and intensity of heat waves. This has also been found in this thesis, in particular identifying a considerable increase in the shape of summer maximum temperatures Central Europe and Northern Italy, areas involved by the growing extension of summer heat waves.

The blended homogenized series are used as input to obtain the homogenized E-OBS, calculated following the techniques documented by Haylock et al. [2008] and Cornes et al. [2018]. The homogenized E-OBS is a reliable representation of the climate of the last decades and can be used for the validation of climate simulations such as those developed in the frame of the High Resolution Model Intercomparison Project (HighResMIP) [Haarsma et al., 2016]. Such validations are fundamental to determine how well the climate models work in reproducing the climate of the past and, by consequence, how reliable is their prediction about the climate of the future.

When checking the trends in the number of winter days with minimum temperatures below the 10$^{\text{th}}$ (1981-2010) percentile, a large overestimation (thus colder trends than observed) is found for several models over Russia, Ukraine and Belarus (Figure 5.3). This indicates that the steep warming of cold extremes in such areas (also observed in Chapter 3) is not well simulated. A similar situation is found for the trends in number of summer days above the 90th percentile in Southern Europe (Figure 5.7). This recalls similar behaviours pointed out by Min et al. [2013] and Sillmann et al. [2013], which reported a clear underestimation of warming trends over various regions of Europe. Such finding suggests that climate projections predict an increase of warm extremes over Southern Europe which is probably lower that what will actually happen.

6.3 Outlook

6.3.1 Possible improvements to the procedure

The Quantile Matching has shown great effectiveness on the ECA&D dataset, fulfilling the requirements that were set at the start of the project. However further ideas for improvements have risen.

Preliminary actions on the dataset would bring relevant benefits to the homogenization procedure. One of this is a powerful quality check, able to detect and eventually remove outliers, repeated values and more kinds of mistakes occurred during the measurement or the transcription of temperature values. The realization of a more sophisticated quality procedure is in progress in the frame of the service C3S311aLot4 (http://surfobs.climate.copernicus.eu) and the project INDECIS (http://http://indecis.eu/). It aims at improving the current one active on ECA&D, which relies on control of the consistency of the data with the statistical characteristics of the series itself, according to the recommendations of the World Meteorological Organization. The new procedure will compare the daily data with the values of neighbouring series, so that the risk of flagging extreme values as outliers is reduced to the lowest. As a result, the (already robust) quantile matching process will deal with only clean and realistic data and be even more effective.

Further improvements on the Quantile Matching procedure are possible: the homogenization procedure is criticized for the influence that is given to the adjustments by references that are clustered in the same area (see Appendix D of Chapter 2). In these situations, the selection of the reference selection can be improved. Although the applied threshold on correlation provides a first check, giving priority to series with similar behaviour, the risk of attributing too large weight to station populated areas is high. Therefore the correlation selection can be coupled with constraints that favour an uniform angular distribution of the references around the inhomogeneous stations. This could be done for example dividing the surrounding region into four quadrants (SouthWest, NorthWest, NorhEast, SouthEast) and selecting an equal number of references in each of them. Such process can provide more solidity to the adjustment calculation, though this will imply a considerable computational cost, since for each station the required time for the reference series selection would be four times longer.

Another critical step in the Quantile Matching procedure is the reiteration of the process. This is performed in order to identify and correct inhomogeneities that were not detectable in the first iteration, when the signal to noise ratio is very low. Ideally the process could be reiterated until no change points are identified by the break detection. However, this choice would trigger a progressive increase of false positives. This is related to the fact that the use of the homogenized series as references for the following iteration determines a process with no memory of the previous stages, likely to diverge from the real original conditions. The choice of limiting to two iteration is purely pragmatic and derives from the evident benefits reached after this stage. Nevertheless, the iterating system could be improved if for each iteration, instead of using the output of previous iteration as references, the original series, split into homogeneous sub-series as of first iteration, are used. Even though this solution would provide a binding constraint to the system, avoiding divergence from the original state, there would still be the need to define when to stop the process, determining a *convergence threshold* (e.g. limited difference between the results of two consecutive iterations). Furthermore such process might bring to an indefinite computational time, so that directly setting a maximum number of iterations would be needed anyway.

6.3.2 Enriching the comparisons of homogenization methods

The comparison of homogenization methods has been the subject of several studies, especially in the last decade [Venema et al., 2013; Domonkos, 2013; Trewin, 2013; Li et al., 2016; Vincent et al., 2018]. This thesis, pursuing on one side the validation of the developed method, in parallel has contributed to the debate on the tools and the methods to use for a thorough comparison. The quality of the developed tools, such as the new kind of benchmark datasets and the use of traditional and innovative metrics, have allowed to obtain interesting results and to confirm the solidity of the Quantile Matching. Nevertheless there is space for a wide spectrum of improvement, that would make a comparison of homogenization methods more challenging and detailed.

One example is the possible refining of the process of benchmark construction with the insertion of random missing data, outliers and data switches between minimum and maximum temperatures. This allows a ”flavours approach”, similar to what adopted in the frame of the INDECIS process, where the author has taken part in a preliminary phase.

Further benefits can be obtained by the use of new metrics, with the aim of testing the robustness and, in particular, the accuracy in reproducing extreme events. For this purpose, metrics as RMSE and PD05 (see Chapter 4 for the definitions) can be calculated restricting the sample to only the values above the 10^{th} or below the 90^{th} percentile of the benchmark allowing to focus the evaluation on the extreme values themselves and not only on the trends.

The procedure of reference selection has shown to be a fundamental aspect in the functioning of a homogenization procedure. Therefore any methods can benefit from the use of more sophisticated reference selections. For this reason the comparison performed in Chapter 4 could be further enriched by the analysis of the combination of homogenization procedures with different reference selection routines. This would give a clearer overview of the capabilities of the homogenization procedures and about the impact of the reference selection.

Finally the homogenization indicator can be upgraded to avoid its sensibility to singularities (e.g. inserting logarithmic behaviours) and can be analysed counting the percentage of series whose value fall in the 0.5-1.5 range, considered the optimal one.

6.3.3 Physical based homogenization

As shown in this thesis, the homogenization of temperature series is a fundamental step in any reliable climatological analysis based on temperature measurements from weather stations. The fact that usually the amplitude of the signal introduced by external interferences is comparable to the climatic signal, makes unreliable any study that skips this phase. The results of any research strongly depends on whether the homogenization step is implemented and the accuracy of the used homogenization method influences the results, as seen in Chapter 4. In particular these variations are observed for the extreme temperature events, whose high sensitivity to inhomogeneities is proven.

The statistical approach developed in this thesis manages extreme values according to their position in their monthly distribution. This allows to distinguish the adjustments for cold and warm events from those for the average. Although the relevant results shown, the inhomogeneities are driven by more complex factors than the simple temperature itself.

In the specific case of a relocation, the difference in temperature between the urban and rural sites is likely to be influenced by the weather conditions. The same incoming solar radiation differently influences the two situations, affecting the amplitude of the Urban Heat Island. In a bright summer day the urban station is likely to measure sensitively higher temperatures compared to the rural station, due to the heat that is re-emitted by asphalt and buildings in the surroundings. On the other side during a cloudy day buildings and roads don't get over-heated by solar radiation and re-emit less than in sunny conditions. Thus the temperature difference between urban and rural station is expected to be lower. Such difference could be even lower in the case of strong wind, which favours the horizontal mixing of the atmosphere.

The sky conditions can be also relevant in the difference of minimum temperatures. A thick cloud cover is expected to absorb part of the thermal long wave radiation emitted by the urban areas (due to e.g. buildings heating) and to partially re-emit downwards, further increasing the temperature difference with the rural station. Night temperature can also be influenced by different humidity conditions or by the presence of snow cover that would further lower the rural temperature.

Other inhomogeneity sources, such as the change of screen, analogously present dependence from the incoming solar radiation and the cloud cover, as it was observed in previous works [Auchmann and Brönnimann, 2012]. Therefore the growing availability of data about these weather variables opens to new cutting-edge evolutions.

The measurements of cloud cover, solar radiation, humidity, wind direction and speed, precipitation and snow depth stored in ECA&D give a precious chance for the development of homogenization procedures that account for physical conditions. Their values could be used to implement a more sophisticated version of Quantile Matching. The sample of temperature measurements can be divided in more sub-samples according, for example, to the cloud cover, as done by Brugnara et al. [2016]. This would allow to calculate adjustments to be applied especially to very cloudy days or to days with clear sky. The same could be done for all the available variables. Such process would create a set of estimations of adjustments from which it would be possible to calculate (for example with a simple mean or median) the final adjustment.

Evidently this kind of approach would require complete and long-lasting measurements for each variable, which unfortunately are not always available in the ECA&D. For this reason it would be difficult to apply it to regions where quality and entirety of data is limited. Anyway, this improvement can be developed as a plug-in, used in case of sufficient data availability. By doing this there would be the possibility to adjust the whole temperature dataset with the normal Quantile Matching and to apply further physycal-based (and probably small) adjustments on the areas where this is feasible. Furthermore, this kind of process will allow to perform adjustments to the temperature measuremente preserving coherence with all the other weather variables, thus conveying to the final product higher quality and reliability.

In the cases where the measurements from other variables are lacking, the information related to the weather condition can be obtained from very high resolution boundary layer models as the Weather Research and Forecasting System (WRF). This takes the reanalysis values of several weather variables (temperature, wind, solar radiation, precipitation, humidity, etc.) as bounday condition and is able to simulate the detailed weather conditions with a spatial resolution of 100 meters and a temporal resolution of 10 minutes.

The use of the results of Weather Research and Forecasting System (WRF) as references for the homogenization was the aim of the MSc internship at KNMI of Jan Biermann [Biermann, 2018], supervised (among the others) by the author of this thesis. Within this project an inhomogeneous series (similarly to what done in Chapter 3) was created concatenating data from Amsterdam Noord between 2008 and 2012 and from Schiphol Airport from 2013 to 2017. WRF has been used to simulate the weather conditions in the area, generating values for both locations. At the same time, the Quantile Matching has been used to reconstruct the temperature in Schiphol in the first period, knowing the temperature in Amsterdam Noord. This has allowed to evaluate the accuracy of the reconstruction of the Urban Heat Island Effect by Quantile Matching, observing that the statistical homogenization procedure reproduces the known values in Schiphol better than the WRF. Such results open to the possibility to join the efforts of Quantile Matching and WRF for the development of a new kind of homogenization that, taking into consideration observed values of all the weather variables and the reanalysis, is able to improve the adjustment according to the weather conditions of the area.

6.3.4 Concluding Remarks

The work displayed in this thesis has been developed with strong links to the previous researches in the field. Several aspects of previous works have been considered and used as inspiration. These have been reinterpreted and combined with innovative solutions, contributing to the creation of a completely automated homogenization procedure. This is able to deal with all the various station density configurations present in the European Climate Assessment ad Dataset and to work on all the kind of inhomogeneities that this dataset presents. These qualities convey the Quantile Matching the prerequisites to become an ideal homogenization methods for other temperature datasets that, as ECA&D, are involved by various configurations and qualities of data.

The developed approach focuses on maintaining as much as possible the characteristics of the original dataset. This makes the homogenized dataset particularly reliable for the analysis of extreme events. Such values are treated with care and adjusted independently from the average. Thanks to this the trend assessments performed on the extreme indices for homogenized temperatures are particularly valuable and trustworthy. These inspections describe a continent that is undergoing a rather dramatic change in its climatic features, with reducing cold extreme events and enhancing warm extreme events.

The discovery that the analyzed climate models don't completely simulate such behaviours is worrying. This calls on the need of further inspecting their accuracy and extend the study on more projects and on wider areas than only Europe. By consequence it is possible to speculate that if they don't correctly reproduce, for example, the frequency and intensity of heatwaves on Southern Europe in the last decades, they are not likely to perform better for the coming decades.

The possible uses of homogenization datasets and the potentialities for other studies prove that the development of homogenization processes for temperature dataset is not a self aimed study. The robustness of the process itself and the reliability of its results allow the improvement of several levels of climatological research and strengthen the validation studies that make use of homogenized series.

Following the aforementioned results, there is strong belief that this study, brings considerable advantages for the scientific research in the field of climate assessment, with particular benefits for the studies on extreme events. This will firstly enhance the solidity and the accuracy of the evidences of climate change and, by consequence, make the communication to society and policy makers more persuasive and effective.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., and Wieringa, J. (2003). Guidance on metadata and homogenization. *Wmo Td*, 1186:53.

Alexander, L. V., Zhang, X., Peterson, T. C., Caesar, J., Gleason, B., Klein Tank, A., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., et al. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111(D5).

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *International Journal of Climatology*, 6(6):661–675.

Alexandersson, H. and Moberg, A. (1997). Homogenization of swedish temperature data. part i: Homogeneity test for linear trends. *International Journal of Climatology*, 17(1):25–34.

Andrade, C., Leite, S., and Santos, J. (2012). Temperature extremes in europe: overview of their driving atmospheric patterns. *Natural Hazards and Earth System Sciences*, 12(5):1671.

Auchmann, R. and Brönnimann, S. (2012). A physics-based correction model for homogenizing sub-daily temperature series. *Journal of Geophysical Research: Atmospheres*, 117(D17).

Begert, M., Schlegel, T., and Kirchhofer, W. (2005). Homogeneous temperature and precipitation series of switzerland from 1864 to 2000. *International Journal of Climatology*, 25(1):65–80.

Bhend, J. and Whetton, P. (2013). Consistency of simulated and observed regional changes in temperature, sea level pressure and precipitation. *Climatic Change*, 118(3-4):799–810.

Biermann, J. (2018). High-resolution reanalysis as reference for homogenization studies - the amsterdam case. *Wageningen University.*

Bindoff, N. L., Stott, P. A., AchutaRao, K. M., Allen, M. R., Gillett, N., Gutzler, D., Hansingo, K., Hegerl, G., Hu, Y., Jain, S., et al. (2013). Detection and attribution of climate change: from global to regional.

Böhm, R., Auer, I., Brunetti, M., Maugeri, M., Nanni, T., and Schöner, W. (2001). Regional temperature variability in the european alps: 1760–1998 from homogenized instrumental time series. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 21(14):1779–1801.

Böhm, R., Jones, P. D., Hiebl, J., Frank, D., Brunetti, M., and Maugeri, M. (2010). The early instrumental warm-bias: a solution for long central european temperature series 1760–2007. *Climatic Change*, 101(1-2):41–67.

Brandsma, T. and Van der Meulen, J. (2008). Thermometer screen intercomparison in de bilt (the netherlands)—part ii: Description and modeling of mean temperature differences and extremes. *International Journal of Climatology*, 28(3):389–400.

Brown, S., Caesar, J., and Ferro, C. A. (2008). Global changes in extreme daily temperature since 1950. *Journal of Geophysical Research: Atmospheres*, 113(D5).

Brugnara, Y., Auchmann, R., Brönnimann, S., Bozzo, A., Berro, D. C., and Mercalli, L. (2016). Trends of mean and extreme temperature indices since 1874 at low-elevation sites in the southern alps. *Journal of Geophysical Research: Atmospheres*, 121(7):3304–3325.

Brunet, M., Saladie, O., Jones, P., Sigro, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D., and Almarza, C. (2006). The development of a new dataset of spanish daily adjusted temperature series (sdats)(1850–2003). *International Journal of Climatology*, 26(13):1777–1802.

Brunetti, M., Maugeri, M., Monti, F., and Nanni, T. (2006). Temperature and precipitation variability in italy in the last two centuries from homogenised instrumental time series. *International Journal of Climatology*, 26(3):345–381.

Caussinus, H. and Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(3):405–425.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., Masina, S., Scoccimarro, E., Materia, S., Bellucci, A., et al. (2019). Global mean climate and main patterns of variability in the cmcc-cm2 coupled model. *Journal of Advances in Modeling Earth Systems*, 11(1):185–209.

Cornes, R. and Jones, P. (2013). How well does the era-interim reanalysis replicate trends in extremes of surface temperature across europe? *Journal of Geophysical Research: Atmospheres*, 118(18):10–262.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J., and Jones, P. D. (2018). An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409.

Cubasch, U., Wuebbles, D., Chen, D., Facchini, M. C., Frame, D., Mahowald, N., and Winther, J.-G. (2013). Introduction. in 'climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change'. *K. Plattner, M. Tignor, SK Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and PM Midgley (Cambridge, UK, and New York: Cambridge University Press, 2013), http://www. climatechange2013. org/images/report/WG1AR5_Chapter01_FINAL. pdf.*

Della-Marta, P. and Wanner, H. (2006). A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate*, 19(17):4179–4197.

Della-Marta, P. M., Haylock, M. R., Luterbacher, J., and Wanner, H. (2007). Doubled length of western european summer heat waves since 1880. *Journal of Geophysical Research: Atmospheres*, 112(D15).

Delvaux, C., Ingels, R., Vrábeĺ, V., Journée, M., and Bertrand, C. (2018). Quality control and homogenization of the belgian historical temperature data. *International Journal of Climatology*.

Dienst, M., Lindén, J., Engström, E., and Esper, J. (2017). Removing the relocation bias from the 155-year haparanda temperature record in northern europe. *International Journal of Climatology*, 37(11):4015–4026.

Diffenbaugh, N. S., Schrer, M., and Ashfaq, M. (2013). Response of snow-dependent hydrologic extremes to continued global warming. *Nature Climate Change*, 3:379–384.

Domonkos, P. (2011). Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theor. Appl. Climatol.*, 105(3-4):455–467.

Domonkos, P. (2013). Efficiencies of inhomogeneity-detection algorithms: comparison of different detection methods and efficiency measures. *Journal of Climatology*, 2013.

Domonkos, P. and Coll, J. (2017). Homogenisation of temperature and precipitation time series with acmant3: method description and efficiency tests. *International Journal of Climatology*, 37(4):1910–1921.

Donadelli, M., Jüppner, M., Riedel, M., and Schlag, C. (2017). Temperature shocks and welfare costs. *Journal of Economic Dynamics and Control*, 82:331–355.

Donat, M., Alexander, L., Yang, H., Durre, I., Vose, R., Dunn, R., Willett, K., Aguilar, E., Brunet, M., Caesar, J., et al. (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The hadex2 dataset. *Journal of Geophysical Research: Atmospheres*, 118(5):2098–2118.

Donat, M. G. and Alexander, L. V. (2012). The shifting probability distribution of global daytime and night-time temperatures. *Geophysical Research Letters*, 39(14).

Easterling, D. R., Peterson, T. C., and Karl, T. R. (1996). On the development and use of homogenized climate datasets. *Journal of climate*, 9(6):1429–1434.

ECA&D Project Team (2012). *European Climate Assessment & Dataset Algorithm The-*

oretical Basis Document (ATBD). De Bilt, NL. version 10.5.

Efthymiadis, D., Goodess, C., and Jones, P. (2011). Trends in mediterranean gridded temperature extremes and large-scale circulation influences. *Natural Hazards and Earth System Sciences*, 11(8):2199.

ETCCDI (2009). Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation. *World Meteorological Organization.*

Fioravanti, G., Piervitali, E., and Desiato, F. (2019). A new homogenized daily data set for temperature variability assessment in italy. *International Journal of Climatology.*

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., et al. (2014). Evaluation of climate models. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 741–866. Cambridge University Press.

Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M., Tank, A. K., and Peterson, T. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate research*, 19(3):193–212.

Giorgi, F. (2006). Climate change hot-spots. *Geophysical research letters*, 33(8).

Gleckler, P. J., Taylor, K. E., and Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6).

Gubler, S., Hunziker, S., Begert, M., Croci-Maspoli, M., Konzelmann, T., Brönnimann, S., Schwierz, C., Oria, C., and Rosas, G. (2017). The influence of station density on climate data homogenization. *International Journal of Climatology*, 37(13):4670–4683.

Guijarro, J. A. (2018). Homogenization of climatic series with climatol. *Reporte técnico, State Meteorological Agency (AEMET), Balearic Islands Office, Spain.*

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J. H., Storch, J.-S. v., Brüggemann, N., Haak, H., and Stössel, A. (2019). Max planck institute earth system model (mpi-esm1. 2) for the high-resolution model intercomparison project (highresmip). *Geoscientific Model Development*, 12(7):3241–3281.

Haarsma, R. J. (2020). Ec-earth3p in progress.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., et al. (2016). High resolution model intercomparison project (highresmip v1. 0) for cmip6. *Geoscientific Model Development*, 9(11):4185–4208.

Hannart, A., Mestre, O., and Naveau, P. (2014). An automatized homogenization procedure via pairwise comparisons with application to argentinean temperature series. *International Journal of Climatology*, 34(13):3528–3545.

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, 48(4).

Hartmann, D., Tank, A., and Rusticucci, M. (2013). Working group i contribution to the ipcc fifth assessment report. *Climatic Change*, pages 31–39.

Haylock, M., Hofstra, N., Tank, A. K., Klok, E., Jones, P., and New, M. (2008). A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research: Atmospheres*, 113(D20).

Jones, P., Lister, D., Osborn, T., Harpham, C., Salmon, M., and Morice, C. (2012). Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*, 117(D5).

Jones, P. D. and Wigley, T. (2010). Estimation of global temperature trends: what's important and what isn't. *Climatic Change*, 100(1):59–69.

Katz, R. W. and Brown, B. G. (1992). Extreme events in a changing climate: variability is more important than averages. *Climatic change*, 21(3):289–302.

Kharin, V. V., Zwiers, F. W., and Zhang, X. (2005). Intercomparison of near-surface temperature and precipitation extremes in amip-2 simulations, reanalyses, and observations. *Journal of Climate*, 18(24):5201–5223.

Kharin, V. V., Zwiers, F. W., Zhang, X., and Hegerl, G. C. (2007). Changes in temperature and precipitation extremes in the ipcc ensemble of global coupled model simulations. *Journal of Climate*, 20(8):1419–1444.

Kiktev, D., Sexton, D. M., Alexander, L., and Folland, C. K. (2003). Comparison of modeled and observed trends in indices of daily climate extremes. *Journal of Climate*, 16(22):3560–3571.

Klein Tank, A. and Können, G. (2003). Trends in indices of daily temperature and precipitation extremes in europe, 1946–99. *Journal of climate*, 16(22):3665–3680.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Milate, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L., and Petrovic, P. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *International Journal of Climatology*, 22(12):1441–1453.

Klok, E. J. and Klein Tank, A. M. G. (2008). Updated and extended European dataset of daily climate observations. *International Journal of Climatology*, 29:1182–1191. doi:10.1002/joc.1779.

Krauskopf, T. and Huth, R. (2019). Temperature trends in europe: comparison of different data sources. *Theoretical and Applied Climatology*, pages 1–12.

Kruger, A. C. and Nxumalo, M. (2017). Surface temperature trends from homogenized time series in south africa: 1931–2015. *International Journal of Climatology*, 37(5):2364–2377.

Kuglitsch, F.-G., Auchmann, R., Bleisch, R., Brönnimann, S., Martius, O., and Stewart, M. (2012). Break detection of annual swiss temperature series. *Journal of Geophysical Research: Atmospheres*, 117(D13).

Kuglitsch, F. G., Toreti, A., Xoplaki, E., Della-Marta, P. M., Luterbacher, J., and Wanner, H. (2009). Homogenization of daily maximum temperature series in the mediterranean. *Journal of Geophysical Research: Atmospheres*, 114(D15).

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., and Rennie, J. (2011). An overview of the global historical climatology network monthly mean temperature data set, version 3. *Journal of Geophysical Research: Atmospheres*, 116(D19).

Li, Z., Cao, L., Zhu, Y., and Yan, Z. (2016). Comparison of two homogenized datasets of daily maximum/mean/minimum temperature in china during 1960–2013. *Journal of Meteorological Research*, 30(1):53–66.

Lindau, R. and Venema, V. (2013). On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records. *Idojaras, QJ Hung. Meteorol. Serv*, 117(1):1–34.

Lindau, R. and Venema, V. (2016). The uncertainty of break positions detected by homogenization algorithms in climate records. *International Journal of Climatology*, 36(2):576–589.

Lorenz, R., Jaeger, E. B., and Seneviratne, S. I. (2010). Persistence of heat waves and its link to soil moisture memory. *Geophysical Research Letters*, 37:L09703. doi:10.1029/2010GL042764.

Mamara, A., Argiriou, A., and Anadranistakis, M. (2014). Detection and correction of inhomogeneities in greek climate temperature series. *International Journal of Climatology*, 34(10):3024–3043.

Maugeri, M., Buffoni, L., Delmonte, B., and Fassina, A. (2002). Daily Milan temperature and pressure series (1763-1998): Completing and homogenising the data. *Climatic Change*, 53:119–149. doi:10.1023/A:1014923027396.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F., Stouffer, R. J., and Taylor, K. E. (2007). The wcrp cmip3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, 88(9):1383–1394.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):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(20):4271–4286.

Menne, M. J. and Williams Jr, C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22(7):1700–1717.

Mestre, O., Gruber, C., Prieur, C., Caussinus, H., and Jourdain, S. (2011). Splidhom: A method for homogenization of daily temperature observations. *Journal of Applied Meteorology and Climatology*, 50(11):2343–2358.

Min, E., Hazeleger, W., Van Oldenborgh, G., and Sterl, A. (2013). Evaluation of trends in high temperature extremes in north-western europe in regional climate models. *Environmental Research Letters*, 8(1):014011.

Mitchell, J., Dzerdzeevskii, B., Flohn, H., Hofmeyr, W., Lamb, H., Rao, K., and Wallén, C. (1966). Climatic change. technical note, no. 79. *World Meteorological Organization: Geneva, Switzerland*, 99.

Morak, S., Hegerl, G., and Kenyon, J. (2011). Detectable regional changes in the number of warm nights. *Geophysical Research Letters*, 38(17).

Nemec, J., Gruber, C., Chimani, B., and Auer, I. (2013). Trends in extreme temperature indices in austria based on a new homogenised dataset. *International Journal of Climatology*, 33(6):1538–1550.

Osadchyi, V., Skrynyk, O., Radchenko, R., and Skrynyk, O. (2018). Homogenization of ukrainian air temperature data. *International Journal of Climatology*, 38(1):497–505.

Pérez-Zanón, N., Sigró, J., Domonkos, P., and Ashcroft, L. (2015). Comparison of homer and acmant homogenization methods using a central pyrenees temperature dataset. *Advances in Science and Research*, 12(1):111–119.

Perkins, S., Alexander, L., and Nairn, J. (2012). Increasing frequency, intensity and duration of observed global heatwaves and warm spells. *Geophysical Research Letters*, 39(20).

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., Toumenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E., Hanssen-Bauer, I., Alexandersson, H., Jones, P. D., and Parker, D. (1998). Homogeneity adjustments of *in situ* atmospheric climate data: A review. *International Journal of Climatology*, 18:1493–1517.

Rahimzadeh, F. and Nassaji Zavareh, M. (2014). Effects of adjustment for non-climatic discontinuities on determination of temperature trends and variability over iran. *Inter-*

national Journal of Climatology, 34(6):2079–2096.

Ren, G.-Y. (2017). Urbanization as a major driver of urban climate change. *Advances in Climate Change Research.*

Roberts, C. D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., and Keeley, S. P. (2018). Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Geoscientific model development*, 11(9):3681–3712.

Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., Jackson, L. C., Kuhlbrodt, T., Mathiot, P., Roberts, C. D., et al. (2019). Description of the resolution hierarchy of the global coupled hadgem3-gc3. 1 model as used in cmip6 highresmip experiments. *Geoscientific Model Development*, 12(12):4999–5028.

Roberts, M. J., Camp, J., Seddon, J., Vidale, P. L., Hodges, K., Vanniere, Benoit and-Mecking, J., Haarsma, R., Bellucci, A., Scoccimarro, E., Caron, L.-P., Chauvin, F., Terray, L., Valcke, S., Moine, M.-P., Putrasahan, D., Robert, C., Senan, R., Zarzycki, C., and Ullrich, P. (2020). Climate model configurations of the ecmwf integrated forecasting system (ecmwf-ifs cycle 43r1) for highresmip. *Submitted.*

Rohde, R., Muller, R., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtele, J., Groom, D., and Wickham, C. (2013). A new estimate of the average earth surface land temperature spanning 1753 to 2011, geoinfor geostat: An overview 1: 1. *of*, 7:2.

Schär, C., Vidale, P. L., Lüthi, D., Frei, C., Häberli, C., Liniger, M. A., and Appenzeller, C. (2004). The role of increasing temperature variability in european summer heatwaves. *Nature*, 427(6972):332.

Scherrer, S. C., Appenzeller, C., Liniger, M. A., and Schär, C. (2005). European temperature distribution changes in observations and climate change scenarios. *Geophysical Research Letters*, 32(19).

Sen, P. K. (1968). Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C. (2006). Land-atmosphere coupling and climate change in Europe. *Nature*, 443:205–209. doi:10.1038/nature05095.

Sillmann, J., Donat, M. G., Fyfe, J. C., and Zwiers, F. W. (2014a). Observed and simulated temperature extremes during the recent warming hiatus. *Environmental Research Letters*, 9(6):064023.

Sillmann, J., Kharin, V., Zhang, X., Zwiers, F., and Bronaugh, D. (2013). Climate extremes indices in the cmip5 multimodel ensemble: Part 1. model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, 118(4):1716–1733.

Sillmann, J., Kharin, V., Zwiers, F., Zhang, X., Bronaugh, D., and Donat, M. (2014b). Evaluating model-simulated variability in temperature extremes using modified percentile indices. *International Journal of Climatology*, 34(11):3304–3311.

Simolo, C., Brunetti, M., Maugeri, M., and Nanni, T. (2011). Evolution of extreme temperatures in a warming climate. *Geophysical research letters*, 38(16).

Simolo, C., Brunetti, M., Maugeri, M., Nanni, T., and Speranza, A. (2010). Understanding climate change–induced variations in daily temperature distributions over italy. *Journal of Geophysical Research: Atmospheres*, 115(D22).

Squintu, A. A., van der Schrier, G., Brugnara, Y., and Klein Tank, A. (2019). Homogenization of daily temperature series in the european climate assessment & dataset. *International Journal of Climatology*, 39(3):1243–1261.

Squintu, A. A., van der Schrier, G., Štěpánek, P., Zahradníček, P., and Klein Tank, A. (2020a). Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theoretical and Applied Climatology*, pages 1–17.

Squintu, A. A., van der Schrier, G., van den Besselaar, E., van der Linden, E., Scoccimarro, E., Roberts, C., and Klein Tank, A. (2020b). Evaluation of trends in extreme temperatures simulated by highresmip models across europe (submitted). *Climate Dynamics*.

Squintu, A. A., van der Schrier, G., van den Besselaar, E. J., Cornes, R. C., and Klein Tank, A. M. (2020c). Building long homogeneous temperature series across europe: a new approach for the blending of neighboring series. *Journal of Applied Meteorology and Climatology*, (2020).

Stainforth, D. A., Chapman, S. C., and Watkins, N. W. (2013). Mapping climate change in european temperature distributions. *Environmental Research Letters*, 8(3):034031.

Štěpánek, P., Zahradníček, P., and Farda, A. (2013). Experiences with data quality control and homogenization of daily records of various meteorological elements in the czech republic in the period 1961–2010. *Időjárás*, 117(1):123–141.

Štěpánek, P., Zahradníček, P., and Skalák, P. (2009). Data quality control and homogenization of air temperature and precipitation series in the area of the czech republic in the period 1961–2007. *Advances in Science and Research*, 3(1):23–26.

Sterl, A., Severijns, C., Dijkstra, H., Hazeleger, W., van Oldenborgh, G. J., van den Broeke, M., Burgers, G., van den Hurk, B., van Leeuwen, P. J., and van Velthoven, P. (2008). When can we expect extremely high surface temperatures? *Geophysical Research Letters*, 35(14).

Syrakova, M. and Stefanova, M. (2009). Homogenization of bulgarian temperature series. *International Journal of Climatology*, 29(12):1835–1849.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498.

Thorne, P., Menne, M., Williams, C., Rennie, J., Lawrimore, J., Vose, R., Peterson, T. C., Durre, I., Davy, R., Esau, I., et al. (2016). Reassessing changes in diurnal temperature range: A new data set and characterization of data biases. *Journal of Geophysical Research: Atmospheres*, 121(10):5115–5137.

Thorne, P. W., Parker, D. E., Christy, J. R., and Mears, C. A. (2005). Uncertainties in climate trends: Lessons from upper-air temperature records. *Bulletin of the American Meteorological Society*, 86(10):1437–1442.

Toreti, A., Kuglitsch, F. G., Xoplaki, E., and Luterbacher, J. (2012). A novel approach for the detection of inhomogeneities affecting climate time series. *Journal of Applied Meteorology and Climatology*, 51(2):317–326.

Trewin, B. (2013). A daily homogenized temperature data set for australia. *International Journal of Climatology*, 33(6):1510–1529.

Trewin, B. and Trevitt, E. (1996). The development of composite temperature records. *International Journal of Climatology*, 16:1227–1242.

Tuomenvirta, H. (2001). Homogeneity adjustments of temperature and precipitation series—finnish and nordic data. *International Journal of Climatology*, 21(4):495–506.

Van den Besselaar, E., Haylock, M., Van der Schrier, G., and Klein Tank, A. (2011). A european daily high-resolution observational gridded data set of sea level pressure. *Journal of Geophysical Research: Atmospheres*, 116(D11).

Van den Besselaar, E., Klein Tank, A., Van der Schrier, G., and Jones, P. (2012). Synoptic messages to extend climate data records. *Journal of Geophysical Research: Atmospheres*, 117(D7).

van Oldenborgh, G. J., Drijfhout, S., Van Ulden, A., Haarsma, R., Sterl, A., Severijns, C., Hazeleger, W., Dijkstra, H., et al. (2009). Western europe is warming much faster than expected. *Climate of the Past*, 5(1):1–12.

Venema, V. K., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., et al. (2013). Benchmarking homogenization algorithms for monthly data. 1552(1):1060–1065.

Vincent, L. A. (1998). A Technique for the Identification of Inhomogeneities in Canadian Temperature Series. *Journal of Climate*, 11:1094–1104.

Vincent, L. A., Milewska, E. J., Wang, X. L., and Hartwell, M. M. (2018). Uncertainty in homogenized daily temperatures and derived indices of extremes illustrated using parallel observations in canada. *International Journal of Climatology*, 38(2):692–707.

Vincent, L. A., Wang, X. L., Milewska, E. J., Wan, H., Yang, F., and Swail, V. (2012). A second generation of homogenized canadian monthly surface air temperature for climate trend analysis. *Journal of Geophysical Research: Atmospheres*, 117(D18).

Vincent, L. A., Zhang, X., Bonsal, B., and Hogg, W. (2002). Homogenization of daily temperatures over canada. *Journal of Climate*, 15(11):1322–1334.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., et al. (2019). Evaluation of cmip6 deck experiments with cnrm-cm6-1. *Journal of Advances in Modeling Earth Systems*.

Vose, R. S., Easterling, D. R., and Gleason, B. (2005). Maximum and minimum temperature trends for the globe: An update through 2004. *Geophysical Research Letters*, 32(23).

Wang, X. L., Chen, H., Wu, Y., Feng, Y., and Pu, Q. (2010). New techniques for the detection and adjustment of shifts in daily precipitation data series. *Journal of Applied Meteorology and Climatology*, 49(12):2416–2436.

Wang, X. L., Wen, Q. H., and Wu, Y. (2007). Penalized maximal t test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, 46(6):916–931.

Williams, C. N., Menne, M. J., and Thorne, P. W. (2012). Benchmarking the performance of pairwise homogenization of surface temperatures in the united states. *Journal of Geophysical Research: Atmospheres*, 117(D5).

World Meteorological Organization (2007). *Manual on the Global Telecommunication System*. Geneva, Ch, wmo-no. 386 edition.

World Meteorological Organization (2011). *Guide on Climatological Practices*. Geneva, Ch, wmo-no. 100 edition.

World Meteorological Organization (2019). *The Global Climate in 2015-2019*. Geneva, Ch, wmo-no. 100 edition.

Yang, P., Ren, G., and Liu, W. (2013). Spatial and temporal characteristics of beijing urban heat island intensity. *Journal of Applied Meteorology and Climatology*, 52(8):1803–1816.

Yosef, Y., Aguilar, E., and Alpert, P. (2018). Detecting and adjusting artificial biases of long-term temperature records in israel. *International Journal of Climatology*, 38(8):3273–3289.

Zhang, X., Hegerl, G., Zwiers, F. W., and Kenyon, J. (2005). Avoiding inhomogeneity in percentile-based indices of temperature extremes. *Journal of Climate*, 18(11):1641–1651.

Ringraziamenti

Final acknowledgements are a tricky step for each PhD thesis. Five years have passed since my first day at KNMI, in a rainy and cold beginning of September. This long period couldn't have been so intense and compelling without the presence of precious colleagues, friends and relatives that helped me getting through it, professionally and personally.

My deepest thanks go to Gerard for the constant support, advice and attention he has given to me, during days and evenings, weeks and weekends, winters and summers. His initiative, his honest opinions, his vast knowledge and his solid "diplomacy" skills have accompanied me, sustained me and formed me as a climate scientist.

A special regard to my office mate Else, for our days spent deep in the data and in the codes, for her incredible capabilities in problem solving and for her unconditional availability to help me facing any inconvenience, including coming to agreements on how to manage shades, windows and fans in our tropical KNMI summer days.

I would like to thank Albert for his help in steering my research projects, for his valuable advices and opinions and for his patience in addressing all my pushy reminders and demanding organizational requests.

What also made these years memorable was the daily life in a department, the RDWD at KNMI, that always made me feel part of a team, where cooperation and mutual help was the basis of our time together. A special mention goes to Ine and Andrew, they were always there for a cup of coffee or a short talk, for a piece of good chocolate or for a good laughter, even in the busiest days. I would also like to thank Richard for the thousand times he has helped me with coding issues.

I particularly thank Corné, my best Dutch teacher and my valuable paranymph, always present next door for any issue or just for a quick chat; Lotte, the best company for coffee breaks and for great dancing nights at conferences (but not only); Marieke, great partner in projects and activities and in Norwegian hiking; last but not least, Irene: our late lunches, our *consplicity* and the mutual support have been fundamental to enjoy good moments and overtake bad times.

Nothing in these years would have been the same without Marina. She has been there all the time, in the good and bad periods, ready for a coffee, for a dinner together, for a good laugh or for deep talks. Even during our fights I have always know that I couldn't rely on anyone like I have relied on her. Thank you for everything.

I particularly thank Pietro for being my first friend when I moved to Utrecht, for our long talks and for designing the wonderful cover of this thesis. Thanks to Lisa for our hundred spontaneous pizzas, for our cappuccinos during the quarantine and for introducing me to the best board game ever, not mentionable here.

A special thought goes to the crew of friends that I have had the luck to meet and hang out with. You are too many to list! Thank you all for our amazing excursions, drinks in the city, dancing nights, board games evenings, Sinterklaas parties and beach days. Don't be disappointed for not being mentioned, just be aware that if you are here today and you got to read these lines, there's a valid reason... and I am deeply thankful for that!

I want to express extreme gratitude for the old friends that made it to today, through the years, no matter the distances. Especially my thanks go to Fede, who was my *rifugio prediletto* in Hamburg and in Ravenna (Filetto) during these years and that will forever be my favourite adventure mate (from acrobatic cycling to creative unvoluntary night conversations). A particular thought for Alessio and Margherita, that through the years have always been up for a chat or a random reunion in a random place. Then yes, I must say thank you to Caco and Guido, even though we barely stand each other and even though our ineptitude is reaching unbearable levels. How not to mention Silvia, we managed to meet each other in Milan, in Parma, in Bologna, in Sardinia, in Pisa, wherever, just because our friendship knows no distance long enough. Finally my gratitude goes to Manu for being always so careful and enthusiast for our friendship and for our long days on the beach.

Friends are precious gems, but those that are present every day, even though they live thousands of kilometers away, are a second family... *un sostegno su cui poter fare riferimento giorno e notte, nella noia e nell'entusiasmo, nei momenti di euforia, di sonno, di fame e di pienezza.* I don't know who would I have been without you. Thank you, Marcu. Thank you, Crispi.

Having a second family is a privilege, even more when there is a first family, *i miei genitori*, like the one I am lucky to have. *In ogni giorno della mia vita, anche nei più difficili, non è mai mancato il vostro abbraccio, il vostro sostegno e il vostro consiglio. Non vi sarò mai grato a sufficienza.* A special thank you for my aunts and uncles, *le mie zie e i miei zii che mi hanno sempre coccolato, supportato e fatto sentire il calore di una famiglia allargata.* Finally all my love to Adele, *dal suo primo giorno, il mio pensiero felice.*

About the author

Antonello Angelo Squintu was born in Sassari, Italy, on the 14th of May, 1989. He attended the "Liceo Scientifico Giovanni Spano" (Scientific High School) between 2003 and 2008, with a final grade of 100/100 (cum laude) and a final thesis about meteorology, with links to Italian literature, philosophy and other disciplines.

Between 2008 and 2012 he has studied General Physics as BSc at University of Pisa, Italy. Here he has graduated (supervised by Prof. Paolo Paolicchi) with a degree of 110/110, discussing a thesis over extra-solar planetary systems detected by the Kepler telescope, focusing on the research of habitable planets.

In 2012 he started the MSc course of Earth System Physics at the University of Bologna, Italy. The final thesis was developed during an internship at the CNR-ISAC, under the supervision of Prof. Michele Brunetti, and focused on the quality check, homogenization and analysis of daily temperature measurements in the region of Trentino Alto Adige, Italy. The graduation took place in 2015 and resulted with a degree of 110/110.

In 2015 he started his PhD at the Royal Netherlands Meteorological Institute, in cooperation with Wageningen University. He has taken part to important European Union funded projects and services such as EUSTACE, UERRA, C3S311a.Lot4 and PRIMAVERA. The work of this PhD project is reported in this thesis.

Peer-reviewed journal publications

Squintu, A. A., van der Schrier, G., Brugnara, Y., and Klein Tank, A. (2019). Homogenization of daily temperature series in the european climate assessment & dataset. *International Journal of Climatology*, 39(3):1243–1261

Squintu, A. A., van der Schrier, G., van den Besselaar, E. J., Cornes, R. C., and Klein Tank, A. M. (2020c). Building long homogeneous temperature series across europe: a new approach for the blending of neighboring series. *Journal of Applied Meteorology and Climatology*, (2020)

Squintu, A. A., van der Schrier, G., Štěpánek, P., Zahradníček, P., and Klein Tank, A. (2020a). Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theoretical and Applied Climatology*, pages 1–17

Squintu, A. A., van der Schrier, G., van den Besselaar, E., van der Linden, E., Scoccimarro, E., Roberts, C., and Klein Tank, A. (2020b). Evaluation of trends in extreme temperatures simulated by highresmip models across europe (submitted). *Climate Dynamics*

Squintu A.A., van der Schrier G., van den Besselaar, E., van der Linden, E., Klein Tank, A. (2020d). Europese temperatuurreeksen: homogenisatie, onzekerheid en effecten voor trends. (accepted) *Meteorologica*

Graduate School Certificate

SENSE

Netherlands Research School for the
Socio-Economic and Natural Sciences of the Environment

DIPLOMA

for specialised PhD training

The Netherlands research school for the
Socio-Economic and Natural Sciences of the Environment
(SENSE) declares that

Antonello Angelo Squintu

born on 14 May 1989 in Sassari, Italy

has successfully fulfilled all requirements of the
educational PhD programme of SENSE.

Wageningen, 16th October 2020

Chair of the SENSE board

Prof. dr. Martin Wassen

The SENSE Director

Prof. dr. Philipp Pattberg

The SENSE Research School has been accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW)

KONINKLIJKE NEDERLANDSE
AKADEMIE VAN WETENSCHAPPEN

The SENSE Research School declares that **Antonello Angelo Squintu** has successfully fulfilled all requirements of the educational PhD programme of SENSE with a work load of 37.3 EC, including the following activities:

SENSE PhD Courses

- Environmental research in context (2017)
- Research in context activity: ‘Taking initiative and organizing a successful “Masterclass Git, GitHub and Markdown in a R-environment”, as a co-production of KNMI and SENSE (26 January 2018 - De Bilt, Netherlands)’

Other PhD and Advanced MSc Courses

- Essential skills in data-intensive research: enabling your research in the life sciences, SURF.nl (2016)
- Swiss Climate Summer School 2016 ‘Coping with uncertainty’ , University of Bern (2016)

External training at a foreign research institute

- ProClimDB Workshop, CzechGlobe, Czech Republic (2016)

Management and Didactic Skills Training

- Tutoring participants to Tularemia (R language course) workshop, University of Utrecht (2016)
- Supervising MSc student Jan Biermann with internship thesis entitled ‘“High-resolution reanalysis as reference for homogenization studies– The Amsterdam case” (2018)

Oral Presentations

- *Homogenization of ECA&D temperature series*, 16th EMS Annual Meeting & 11th European Conference on Applied Climatology, 12-16 September 2016, Trieste Italy
- *Quality Check and Homogenization of ECA&D temperature data-set*, 11th EUMETNET Data Management Workshop, 18-20 October 2017, Zagreb, Croatia
- *Homogenization of ECA&D temperature dataset and comparison of methods*, European Geophysical Union General Assembly, 8-13 April 2018, Vienna, Austria
- *Homogenization of ECA&D temperature dataset – Constructing long series*, 18th EMS Annual Meeting & 13th European Conference on Applied Climatology, 13 September 20-19, Budapest, Hungary

SENSE coordinator PhD education

Dr. ir. Peter Vermeulen

This research received funding from EU FP7Collaborative Project UERRA (Uncertainties in Ensembles of Regional ReAnanalysis), grant agreement 607193, the EU H2020 EUSTACE Project, grant agreement 640171, the Copernicus Climate Change Service through contract C3S 311a.Lot4 and the PRIMAVERA project, funded by the European Union's Horizon 2020 programme, grant agreement 641727.

This is thesis has been printed by ProefschriftMaken.

Financial support from Wageningen University and KNMI for printing this thesis is gratefully acknowledged.

Cover design by Pietro Lodi