Statistical Quality Assurance and Peer Review in Primary Data Publication

Dissertation

zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
André Düsterhus
aus
Salzkotten

Bonn, Oktober 2012

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Andreas Hense
2. Gutachter: Prof. Dr. Clemens Simmer
Tag der Promotion: 7. Januar 2013
Erscheinungsjahr: 2013

Abstract

Publication of results is an elementary part of scientific work. Started in in 17th century with the traditional scientific publications several new forms emerged in recent years, triggered by the digital revolution and focused on increasing the data availability. These changes offer new chances for scientists and science in general, but also imply risks, which might compromise this established institution in science.

Data can be seen as an essential foundation of science. Therefore, it is important to think about new ways to distribute scientific data between the scientists, which are generated by the advent of the world wide web. One way, the primary data publications, aims at the publication of raw data and their metadata and seeks to be comparable to the traditional forms of publishing manuscripts.

This thesis will present the possible ways to publish data, which are typical in meteorological and climatological sciences. Additionally, it shows, how a publication process of the primary data itself can be included into the traditional scientific working scheme. Thereby, it will especially focus on the development of an effective quality assurance of these publications. A fundamental part of this will be general quality tests, which are needed to obtain estimations on the quality of datasets. These quality checks are characterised by its parameter-driven flexibility and can be applied on a lot of different types of datasets. Some of these checks, like a newly developed histogram test and a bayesian change point detection, are described and undergo some sensitivity tests. In addition, the possible automatisation and application to meteorological and climatological datasets of the tests will be investigated. All these developments will be discussed under the aspects of a usability in data publications, an effective quality assurance system and the possibility to generate a peer review procedure on data publications.

Zusammenfassung

Die Veröffentlichung von Resultaten ist ein elementarer Teil des wissenschaftlichen Arbeitens. Angefangen im 17. Jahrhundert mit den traditionellen wissenschaftlichen Publikationen, entstanden in den vergangenen Jahren neue Formen, die sich auf das Zugänglichmachen von Daten fokussieren und durch die digitale Revolution gefördert wurden. Diese Veränderungen beinhalten neue Möglichkeiten für Wissenschaftler und die Wissenschaft im Allgemeinen, aber schliessen ebenso Risiken ein, die möglicherweise eine etablierte Institution der Wissenschaft gefährden.

Daten können als eine wichtige Grundlage der Wissenschaft angesehen werden. Darum ist es von großer Bedeutung, über die neuen Wege des Datenaustausches zwischen Wissenschaftlern nachzudenken, die durch das Aufkommen des World Wide Webs entstanden sind. Ein möglicher Weg ist die Primärdaten Publikation, die versucht, Publikationen von Roh- und ihren Metadaten vergleichbar zu der traditionellen form der Veröffentlichung von Manuskripten zu gestalten.

Diese Arbeit stellt die möglichen Wege zur Publikation von Daten, welche typisch sind in der Meteorologie und Klimatologie, vor und untersucht, wie sich ihr Publikationsprozess in einen traditionellen wissenschaftlichen Arbeitsablauf einfügt. Dabei wird besonders auf eine effektive Qualitätssicherung dieser Publikationen eingegangen. Ein grundlegender Bestandteil sind dabei allgemeine Qualitätstests, die benötigt werden, um Abschätzungen über die Qualität der Datensätze zu erhalten. Diese Qualitätstests sind charakterisiert durch ihre parametergetriebene Flexibilität und können auf viele unterschiedliche Typen von Daten angewendet werden. Einige dieser Tests, wie ein neuer Histogramm Test und eine Bayessche Bruchpunktsdetektion, werden beschrieben und mehreren Sensitivitätstests unterzogen. Es wird ebenfalls eine mögliche Automatisierung und eine Anwendbarkeit der Tests auf meteorologische und klimatologische Datensätze untersucht. Alle diese Entwicklungen werden unter den Gesichtspunkten einer Nutzbarkeit in Datenpublikationen, eines effektiven Qualitästssicherungssystems und der Möglichkeit der Konstruktion eines Peer Review Prozesses von Datenpublikationen diskutiert.

Contents

1 Introduction

Data are a central foundation of science. Most scientists depend on data as a basis for their theories or as a result of their research. As a consequence, well generated and documented datasets are a highly desired commodity among scientists although some scientists are hardly convinceable to share them. Even when the willingness is there, the effort to generate them is very high, what results in a low number of publicly available of these high standard datasets.

Scientific publications are for several hundred years the basic form to communicate and share results of research. They include a documentation of the performed steps, what leads to the generation of new knowledge, and build a basic platform for scientific credit for the authors and their career chances. High standards have been developed, which ensure the quality of published manuscripts. One of them is peer review, which can be seen as the "gold standard of quality control" in the modern scientific world (Drott [2007]).

In the last two decades this well established system has been enhanced and new forms of publications have been developed. Some of these publications tried to support the exchange of data between scientists, which was mainly driven by the arise of the world wide web. One of these new forms is primary data publication, which tries to bring the raw, documented data of research to the standard of the existing traditional publications. An open point of this new form is the assurance of the quality, not only from the technical point of view, but also concerning its content.

This thesis tries to connect the traditional and the new forms of publications and develops procedures and methods, that help to bring both types on the same standard. Therefore, the point of view of the scientist is needed in two different ways. On the one hand there is the data author, who wants to have an effective system, which minimises the effort to publish data. On the other hand there is the data reuser, who wants to have well documented and quality assured data. These two views contradict themselves, what makes it complicate to find ways which balance them and support science in general.

Environmental sciences work with an immense variety of forms and types of data and meteorological and climatological sciences are a good representative of this. Data may be generated at laboratories, in field campaigns or by simulations, but the methods, which assure their quality have to work on all types equally well. The natural way to achieve this is statistics, which is also chosen in this thesis. It will be shown, how general statistical quality checks can be embedded into a concept, which allows to generate effective procedures for the quality assurance. It will also be shown, how this fits into a traditional working scheme of scientists and which types of publications cover which part of the scientific process.

A focus will be set on the development of general quality checks. Those are parameter-driven statistical checks, which allow to estimate the quality of an immense variety of datasets. With the help of automatisation techniques, it is possible to enhance their effectiveness as well as to show ways, how a peer review of data might look like.

The outline of this thesis starts in chapter 2 with a look at the traditional publications and their development and tasks since the 17th century. It also shows the general concept of quality assurance and their implementation for the different types of publication. In addition, software concepts will be presented, which assist the data author in performing the quality assurance on his/her primary data publication. One important part of this will be general quality checks. Examples for these are presented in chapter 3. Mainly two different types are shown, the histogram test and a change point detection, which allows to analyse very different types of datasets. To show the abilities of these methods and their combination, some applications to different datasets will be shown in chapter 4. This is followed by an extensive discussion of the shown procedures and methods in chapter 5. The latter also covers some remarks, on how a peer review of data may look like. In the last chapter 6, a brief conclusion will be given and some remarks on possible further investigations in the future are made.

2 Publication of environmental data

Nowadays, scientific publications are the basis of science. This statement is also valid for environmental science, like meteorology and climatology, which depend strongly on data. At the beginning of this chapter, in section 2.1, the importance and types of meteorological data are presented. In the following section 2.2, history and development of traditional publications are shown. The focus is especially set on the new types of publications, which are the main topic of this thesis. A further look is taken on data publication, which will be introduced in section 2.3. There are two different kinds of new forms of publications, which will be described: the data journal and primary data publication. Important factors for these publications are the quality assurance processes, which are explained and characterised in section 2.4. The data publication process, which is described in section 2.5, was introduced at the World Data Center for Climate (WDC-C). There, the developed software for quality assurance on metadata will be explained in more detail. In a last section 2.6 the scientific quality assurance system for primary data publication will be presented. This includes software solutions and concepts of documentation.

2.1 Meteorological data

Meteorological datasets can be characterised as heterogeneous, from which follows that they are a good representative for environmental data in general. Since meteorological and climatological datasets will be the focus of this thesis, they are introduced at this point in detail. The section starts with a motivation of the importance of data in the meteorological and climatological sciences in section 2.1.1. In a next step, an overview of datasets used in meteorology is given. In section 2.1.2, the datasets will be divided into different data classes and their characterisation will be briefly discussed. In addition, the connection between meteorological and environmental datasets in general are mentioned there. As a consequence, the results of this thesis can be transferred to other sciences than meteorology and climatology.

2.1.1 Importance of data for meteorological and climatological sciences

Data play an important role in meteorology and climatology. Scientists in general depend on data since knowledge is "[...] gained from empirical and modelled data and observations" (Costello [2009]). This is also phrased by Klump et al. [2006] with: "Scientific knowledge is communicated through scientific literature. Knowledge is ultimately derived from data." Therefore it is important, that a broader public is able to get the data and to advance their knowledge for the society.

In the standard working procedures, data are of great importance for scientists working in the field of meteorology and climatology. A scientist creates data by performing experiments. These can happen outdoors in the field, in laboratories or by computer simulations (Overpeck et al. [2011]). They also use data for analysis and create new datasets by transferring original data into new, for their field of research more usable, forms. At the end they use the data for visualisation, to underline their scientific arguments and to purpose new directions for the research of the future.

Nevertheless, data is not only used by scientists to generate new scientific findings. It is also used by others to control the findings, to check whether concluded results are true or questionable. New findings and methods offer new ways to reanalyse existing datasets and to conclude new or dismiss former results.

Another factor, which makes data so important for scientists, is their uniqueness. It is not possible to recreate a dataset exactly once it is lost. This is obvious for observations in nature, where the surroundings of the measuring instruments change continuously. Not that obvious it is for measurements in laboratories. Even performed under controlled environmental conditions, they include uncertainties, which can not be reproduced exactly. The same is valid for computational models. Those results depend not only on the model, input parameters and the used machine for the calculations. Even if all these factors are equal, the results can vary from run to run. Reasons can be found in bad programming, introduced stochastic elements or in the enhancement of efficiency by giving up bit-reproducibility as a basic request to the calculations (Palmer [2012]).

As a consequence, archiving and access to data are essential elements to allow successful scientific research.

2.1.2 Data classes in meteorology

Meteorological data consist mainly of one- or multidimensional time series. The acquiring of these datasets can be divided into two basic parts. The first are data produced by models. Models, which simulate weather or climate, play an integral part in meteorological and climatological science (Overpeck et al. [2011]). A main characteristic of the resulting data are their structure. The majority of the model data is regularly structured. In the spatial domain most of the data is available on regular grids. In addition, in the temporal domain a lot of datasets resulting from simulation can be found with a regular time step. These regularities have advantages for the analysis of the data. A disadvantage can be the multidimensionality of the data. Model runs deliver several variables for every dimension in space and time (Meehl et al. [2007]). Also Monte Carlo simulations, like ensembles, are a common tool to estimate the uncertainties of models. In these, the initial and boundary conditions are varied to create different realisations of the model (Lorenz [1963], Molteni et al. [1996]). Therefore, six dimensional datasets are no more unusual. These dimensions are mostly: three dimensions in space, one in time, one for the different meteorological parameter and the sixth for every realisation of the model. This immense amount of data brings new challenges – not only concerning the analysis of the datasets, but also concerning storing, visualisation and documentation.

The second class of data in meteorology are the observational data. These data have to be further divided into two subclasses. On the one hand there are station data of permanent networks, like for example at the national weather services. These networks are also characterised by regularity in space and time for longer time spans. On the other hand there are measurement campaigns. The data resulting from these campaigns have to be expected as irregular. Additional challenges are the used instruments. In stationary networks the use of generally used, well tested and calibrated instruments can be expected. In field campaigns, especially those, which are done for the purpose of research, instruments in an experimental phase of development are common. This makes it especially problematic to quality assure the resulting datasets.

These differences in meteorological data classes require general and large-scale approaches to handle, store and analyse the data. Such a requirement is not only a burden, but also an advantage, since it allows to generalise the methods to other fields of science. Therefore, meteorological data deliver a good test field as a representative of environmental data. Environmental data include, beside the data on the atmosphere, also the ones on the solid earth, ocean and the biosphere. All these sciences use similar methods and are characterised by using time series as a common form of result for their measurement. This allows to transfer the main results of this thesis to the above mentioned fields and their applications.

2.2 Traditional publication process

Publications on paper, in the following called 'traditional publications', are the basis of scientific research today. Their role and development will be examined in this section. Hence, section 2.2.1 gives a short historical overview of scientific publications. It shows the functionality of this tool within the scientific process. Afterwards, the rank of these publications within a scientific working scheme is shown by means of an idealised traditional scheme in section 2.2.2. This allows to illustrate, which working steps lead to a publication. Followed by this, the new developments in the last years are described in section 2.2.3. At the end the importance for the scientists of such a publication is further explained in section 2.2.4.

2.2.1 Development of scientific publications

The first scientific journals, similar to the ones today, emerged in the middle of the 17th century. The first was named "Journal des Sçavans". It included several scientific essays and was introduced in January 1665 by Denis de Sallo (Brown [1972], Benos et al. [2007]). More important for scientific work today was the publication of the Journal "Philosophical Transactions of the Royal Society" two months later (Philosophical Transaction Staff [1665], Kronick [1990]). It was edited by Henry Oldenburg, who is often credited to have introduced the first modern scientific journal (Pfeiffenberger and Carlson [2011], Kronick [1990]). At the beginning, Oldenburg as the editor was responsible for the content of the journal (Spier [2002]). His idea was to invent a register for innovations in science (Cassella and Calvi [2010]) and to offer scientists of these days the possibility to earn credit for their scientific results, without publishing a book (Pfeiffenberger and Carlson [2011]). Up to these days these scientific journals fulfil four main functions: First, the registration of whom was the first of a finding. The second function is the certification of this claim. The third is to bring awareness of new developments in science to other scientists. The archiving of the findings for preservation can be seen as a fourth (Cassella and Calvi [2010]).

A main part of the scientific process today, the peer review system (see also section 2.2.2), was introduced at the Philosophical Transactions around a hundred years later.[1] It was adopted from the journal "Medical Essays and Observations" which was published by the Royal Society of Edinburgh in 1731 (Benos et al. [2007], Spier [2002]). The influence and the forms of peer review depended from then on on several limiting conditions. The most important one is the space available for publications in the respective journal. Editors, which have to fill their journals to get commercial success would not set the requirements for accepting papers too high. Here, also the number of submitted papers play an important role. This number depends again on the number of people working in science, who want to publish their findings (Burnham [1990]). In addition, the technical progress over time plays an important role. An example is the possibility to multiply the manuscripts (Spier [2002]). A last factor to mention here is the willingness of an editor to give parts of the control of what will be published out of his/her hand. An example for this is one of the leading journals in science today: Science (Fersht [2009], Deutsche Forschungsgemeinschaft [1998]). This journal did not start to adopt some kind of peer review before the 1930s (Knoll [1990], Burnham [1990], Spier [2002]).

With these functionalities and their long stretched history, scientific publications are a foundation of science today.

[1] In a report of the Royal Society (Boulton et al. [2012]), which was published in June 2012, it was claimed, that Henry Oldenburg also invented the peer review at the Philosophical Transactions. It was not possible to verify this claim by other literature.

2.2.2 Traditional scientific working scheme

In recent years the scientific publication process has changed dramatically. New forms, mainly electronic, have emerged and changed the style of how scientists work today (Lancaster [1995]). Therefore, a development of a traditional working scheme will be the next step. It shows how publications can be generated by scientists. This working scheme will be used afterwards to show and explain these new forms of publications. Publishing in general is at the end of a line of several working steps. Its functionality is to document and to present the results of the research and the work behind it. How the scientists get to these results is not only a practical, but also a philosophical question. This section structures the possible steps to a publication and illustrates these steps with an idealised scientific working scheme.

In the following it is assumed, that the scientist who performs the work is doing something like an experiment, which produces data to achieve results. This is a valid assumption for a lot of scientific publications, but does not necessarily mean that someone performs a measurement with an instrument in the field. Alternatively, it can be a model experiment done with computers or, to take an extreme example, a thought experiment, which might for example generate a mind map. To get a scientific result from the experiment, two basically different forms of causality are described in the literature: hypothetico-deductive and inductive (Williamson [2005], p. 118ff). The first bases on predictions by hypotheses, which should be tested by an experiment (see also Popper [1934]). The second, the inductive approach, uses a large number of observations to create a theory of the underlying mechanism, which created these observations (see also Bacon [1620]). Of course, mixed versions of both are also possible. These are intensively discussed by Williamson [2005] (p. 148ff).

To demonstrate the working process of a scientist, a flow chart is created, which is sketched in section 2.1. It demonstrates the basic steps from the idea of the experiment to the publication of a paper. To prevent a restriction to one of the views mentioned above, the used idealised model of a scientific working scheme consists of two phases of theory. The first, the pre-experimental theory, covers together with the prediction the hypothetico-deductive view. In this theory, the scientists construct a hypothesis, which is able to make a prediction. The aim of the experiment is now to back or falsify this prediction. In order to be able to perform this experiment, it has to be designed by the scientist. After the experiment is done, the data is collected and stored. In a next step follows the second theory, the post-experimental theory, which for example covers the statistical analysis of the dataset. Afterwards, a step follows, which is not directly done by the scientist himself/herself, since on the analysis a peer review is performed at a journal. After passing this, the work can be published as a paper.

The process varies from case to case and depends on the types of experiment and analysis. Nevertheless, the most experimental driven scientific processes should be able to cover with this type of idealised working scheme. It will also be used in section 2.2.3 to demonstrate how new developments in scientific publications can be integrated and compared to the traditional working scheme.

This will now be illustrated in more detail with an example. Assumed that a researcher wants to know if a city produces an urban heat island effect (Giridharan et al. [2005], Oke [1973], Oke [1982]). First of all there is the hypothesis, which can be formulated negatively or positively. In this example, a negative formulation is used: "There exists no heat island effect in the specified city". Followed is this step by the prediction. It could be assumed, that it is not possible to measure a temperature difference between temperature measurement stations inside and outside the city over a defined length of measurement time. After this prediction is established, the experiment can be designed. This 'experimental design' covers for example the number and placing of measurement stations, the time when they were build up, the type of the measurement devices and their calibration procedures and a lot more. When this is finished and the resources are available, the experiment is ready to start. The devices will or will not be placed like planed and the measurement takes place. Outcomes will be stored as datasets, quality controlled and

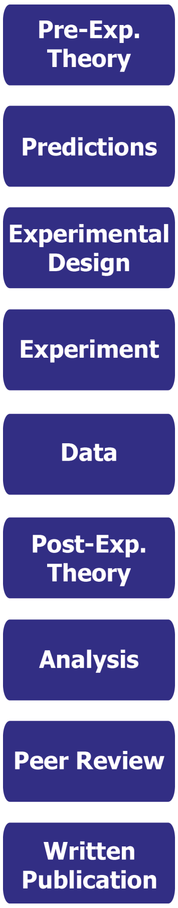


Figure 2.1: An idealised traditional scientific working sheme, which shows the steps from a pre-experimental theory to a written publication.

enhanced with metadata. After this, the statistical analysis, like fitting regression lines of the observed data, is carried out. A statistical hypothesis test might be performed on whether the difference of the measured temperature between the stations in- and outside of the city is significant or not. This simple example illustrates, that a scientific work can be well covered by this scheme.

When the authors decide, that they want to share their findings with other scientists, they may create a publication. The arising question is, what a publication is or should be. Parsons et al. [2010] described it as follows: "A scientific publication is fundamentally an argument consisting of a set of ideas and expectations supported by observations and calculations that serve as evidence of its veracity."

In the traditional publication scheme follows the next step: the analysis. This phase should produce a manuscript, which explains all steps performed before. It also discusses and concludes the theory, the experiment and its results. In a typical structure of a paper, all steps mentioned above are described and explained. After finishing the manuscript the scientist as an author submits it to the journal.

Journals handle the incoming manuscripts differently, but a common practice is the peer review process. Therein an editor makes the first basic review on formalities. After this is passed, he/she choses one, mostly two or more reviewers and transfers the manuscript to them (Campanario [1998a]). Those reviewers perform a review and file a report to the editor. If the editor decides, based on the reports by the reviewers, that the manuscript is publishable without corrections or not publishable with any corrections, the peer review process is finished at this stage (Gura [2002]). In the case, that it has to be corrected, for example when the reviewers have questions, the editor transfers the reports of the reviewers to the authors. They are obliged to comment on the reviews and if necessary to correct the manuscript. The process of reviewing, correcting and editor decision may iterate, until a final decision is made by the editor (Benos et al. [2007], Hargens [1990]).

There are a lot of varieties of this peer review process, which are implemented differently by the journals (Lawrence et al. [2011]). For example, the process is called blind or anonymous, if the author does not get the information about who the reviewers are (Weller [1990]). It is called double-blind, if additionally the reviewers are not getting any information, which might disclose the identity of the author (McNutt et al. [1990], Campanario [1998b]).

If the decision of the editor is positive, the manuscript will be published and becomes a written publication, also called paper. The journal publishes this paper on its platforms and other scientists are now able to get the document and inform themselves about the performed experiment.

The here developed structure is flexible and can therefore be used for a wide variety of scientific processes, which will be further discussed in section 5.3.1.1. Nevertheless, the working scheme experienced several additions in the recent years, of which some will be explained in the following.

2.2.3 New developments in recent years

Since the advent of the world wide web and its possibilities, new forms of publications and additions to the traditional publication process emerged. These changes to scientific publication and its processes went far beyond of just making existing paper journals available in electronic form (Lancaster [1995]).

Electronic publishing started in some forms in the 1960s, but was mainly a form of distributing already existing paper publications. The first complete electronic journals started at the end of the 1970s (Lancaster [1995]). Afterwards, the world wide web boosted the number of electronic journals so that nowadays nearly every journal is (also) available in electronic form (Renear and Palmer [2009]).

In addition, the traditional publication process itself was reformed. For example, some changes to the peer review process were developed. The so called 'open review' allows scientists, who are not formally assigned as reviewers, to comment on the manuscript (Benos et al. [2007]). The additional commentators could either be a closed group or participants of an open discussion. The latter was institutionalised for example by Copernicus (Pöschl [2010]). They introduced discussion papers, which collect these comments and publish them alongside the original publication (ESSD [2012b]). This is very similar to pre-publication of manuscripts in the world wide web. A platform for pre-publications is for example ArXiv and was founded in 1991 (Cassella and Calvi [2010]). It collects manuscripts, which have not passed a peer review process yet and offers them for everyone interested (Warner [2005]).

This system works, because open access, which allows the reader to access articles without paying a fee, becomes a common element in a modernised scientific working process. This dramatically changed the way how scientists access articles and allows them to get amounts of information on the works of others like never before (Meehl et al. [2007], Cassella and Calvi [2010]). There are also new journals established, called overlay journals, which collect open accessible pre-publications of their field, perform their own peer review process on it and publish it as a journal on their own (Cassella and Calvi [2010], Warner [2005]).

Processes like a reformed peer review process or pre-publications, can be comprehended as a quality assurance of the manuscript. This quality assurance is located outside the traditional peer review process and enhances it. These measures coexist with the traditional form for the most of the published publications.

Newer developments aim to concentrate on publishing datasets and their explanation. The so called 'data journals' will be explained in more detail in section 2.3.3. These new forms have to be integrated into the working scheme, which will be done in section 2.3.4, alongside the primary data publication. Before this will be introduced, the importance of publications for scientists should be explained.

2.2.4 Importance of publications for scientists

Publications are not only in the interest of the journals, but of course also of the scientists themselves. It is of great importance for them to publish their findings in well-established journals (Kinne [1988]). Those publications are used to judge the scientists' abilities and are therefore a very important component to build up their career (Campanario [1998a]). When scientists publish, it is not only the content and form of the publication, which is important for them, but also the platform, where they do it. This is for example commented by Casadevall and Fang [2009] with: "One of the fascinating aspects of the sociology of science is that scientists prefer to publish in journals that present the greatest hurdles, which translate into scientific prestige." Therefore, they tend to choose journals with a rigid peer review system and the highest impact factors. This leads to a fifth function of a scientific journal, apart of registration, certification, awareness and archiving: reward to the author (Warner [2005]).

The career perspectives of scientists are not the only important factor, since publishing helps also them to enhance their abilities. For example, they can benefit of good reviewers, who help them to find mistakes in their work or clarify their findings (Casati et al. [2010]).

The other part of importance for scientists is getting knowledge of the works performed by other scientists. It enables them to build their work upon others findings. They are also able to leave some of the responsibility for the cited findings to other scientists (Costello [2009]). This makes it simpler for the scientists to concentrate on their own work. As a consequence, they do not have to put too much efforts in results, which are already found by others, but are necessary as a basis for their own research.

All these arguments show, that scientists accept high effort for publishing, if they see, that they profit from it. This should be kept in mind, when high requirements for publishing data are requested from the scientists. The procedures for this data publication are explained in the upcoming section.

2.3 Data publication

Apart from describing research in written form in journals, the publication of the foundation of the research plays an important part in science today. In this section, the publication of primary data and its actual developments will be described. It starts in section 2.3.1 with an explanation, why primary data publication is so important for the communication in science and for the scientist themselves. In the following section 2.3.2 the actual situation of data publishing will be described. It shows especially the role and ideas of funding agencies in this process. Afterwards, data journals, which are one new emerging form of publication in recent years, will be presented in section 2.3.3. In a last section 2.3.4, the other new form of publishing data, the primary data publication, is discussed. In addition, at this point both types are included into the scientific working scheme, which was introduced in section 2.2.2.

2.3.1 Importance of primary data publication

An essential question on primary data publication is the following: Why should a scientist publish his/her research data? The simplest answer is, that the general public or at least other scientists in the field, can work with the data. The general importance of data for the scientists was already emphasised in section 2.1.1.

Especially smaller research projects depend on a well-established data publication infrastructure. First of all, the money by the funding agencies for these projects is limited, so that they need to rely on sources of data by others to perform their own research. This would be simpler to perform, if these projects had the opportunity to cite data, which they are able to trust. Trusting the data is essential for seeing an alternative in 'using data of others' in contrast to 'producing everything on their own'. Secondly, when it comes to archiving the data, which the scientists produced in these smaller projects, it is common, that their budget for making the data available in useful forms to others is also limited (Heidorn [2008]). This may lead to the problem, that the data is effectively lost after the end of the project, in case that no archiving is performed at all (Klump et al. [2006]).

For all scientists there is additionally the problem of long time archiving. When scientists decide, that data might not be useful for them anymore, they might delete it (Klump et al. [2006]). Nevertheless, this does not mean, that this data is useless for other scientists (Heidorn [2008], Guralnick et al. [2009], AGU Council [2009]). It is also important that data, which is used to postulate theories is saved for the future, because else the fundamental scientific "principle of replicability" (Heidorn [2008]) might be threatened (Strebel et al. [1998]).

Therefore, it is essential for effective and reliable science to preserve data, and make them accessible for others. This is not a new discovery, but rather a basic scientific procedure (Schofield et al. [2009], AGU Council [2009], Kinne [1988]).

Publishing data, even if the infrastructure is available and simple to use for the scientist, still needs a lot of effort. What are his/her benefits for doing it? It can and should be peer-recognition (Costello [2009], Toronto International Data Release Workshop Authors [2009]), which is traditionally expressed in science by citations of the work by others. To generate those citations, the datasets themselves have to be citable. The natural way to achieve this is to use data publication and bring this to a similar standard like journals already have (Costello [2009], Klump et al. [2006]). Obviously, this needs more effort by the publishing scientists than simply putting the raw data on a server and connect it to the internet (Strebel et al. [1998]).

Apart from this, concerns that published datasets will not be cited like articles in journals, are still present (Costello [2009]). Therefore, the question arises if this is really a threat and what reasons exist to believe, that citing datasets in traditional publications will be a common tool in the future. If a dataset is cited at all, depends of course highly on its individual relevancy, quality and the documentation. Additionally, there is the necessity of a general change in the "sociology of science" (Heidorn [2008]). This can be supported for example by the pressure of funding agencies, which want more scientific results to be generated from their money spent on costly experiments (Toronto International Data Release Workshop Authors [2009]). For example, the directorate for geosciences of the National Science Foundation (NSF) calls for establishing data citation "[...] as the rule rather than the exception" (Killeen [2012]).

Should more scientists use and cite existing datasets, prepared by others, investments in those experiments would deliver more scientific output. There is also a study, performed in astronomy literature, that indicates that authors, who cite datasets, earn a higher citation rate for their own paper (Henneken and Accomazzi [2011]). These examples show, that a change in sociology of scientists might be possible in the future. The importance of data availability and publishing shown in this section leads to the conclusion that special forms of publications, which focus solely on data, might be justifiable. In the next section, the actual situation in this field will be further described.

2.3.2 Situation of data publishing today

By speaking of the future and possibilities of primary data publication, it is important to to take a look at the current situation. The problem, that too little amounts of data are well archived and accessible for others is recognised. Research organisations like the American Geophysical Union (AGU) ask their members to making data available (AGU Council [2009]). The funding agency Deutsche Forschungsgemeinschaft (DFG) recommends in their "Proposals for Safeguarding Good Scientific Practice": "Primary data as the basis for publications shall be securely stored for ten years in a durable form in the institution of their origin." (Deutsche Forschungsgemeinschaft [1998], p. 55). A current development at funding agencies can be observed at the National Science Foundation (NSF) in the United States. In January 2011, they added to their recommendations, that for granting a proposal a data management plan has to be included. This covers "Plans for data management and sharing of the products of research, including preservation, documentation, and sharing of data, samples, physical collections, curriculum materials and other related research and education products [...]" (National Science Foundation [2011], chapter II.C.2.d.i). Recently, also the Research Councils of the United Kingdom emphasised in their modification of their "Policies on Access to Research Outputs" in July 2012, that researchers, who are funded by these institutions, have to include "[...]a statement on how the underlying research materials – such as data, samples or models – can be accessed." in their publications (Research Councils UK [2012]).

Apart from these recommendations there are strong reasons, why data is actually not shared. For example, governmental interests, like making commercial receipts from the financed research, are at stake (Overpeck et al. [2011]). Other problems can be found in the scientists themselves. Having well prepared data is a big advantage for them. They may have problems to share the data with their probable opponents, fearing to compromise their career chances (Schofield et al. [2009]). As a result, even big campaigns, like the International Polar Year 2007/2008, struggle with the lack of data availability (Carlson [2011]).

To overcome these difficulties, there are many initiatives, which are mostly based on rewarding or force to store and publish data. Many journals for example ask their authors to make data available on request (Hrynaszkiewicz et al. [2010]). The same is valid for scientists, who are funded by the governments of the United Kingdom or the United States (Jubb [2012]). A major contribution to tackle these problems are the fundings of the world data centres, which will be explained in more detail in section 2.5.1.

To get more insight into the current situation a look at a poll of the journal Science will be taken in the following (Science Staff [2011]). The survey on 1700 scientists indicated, that around eighty percent of the respondents think, that they were not funded enough to curate their data. The same poll shows, that half of the responding scientists save their research data in their own lab, which is far from ideal in terms of long time archiving the data. Another interesting point of this survey is, that half of the respondents use data from archival databases only rarely for their own research.

This all leads to the view, that the problems around data archiving, using and publishing in science are recognised and addressed by major representers of the scientific communities. Nevertheless, it shows also that there is still much work to do in this field in times ahead. Examples for such work are data journals, which will be presented in the upcoming section.

2.3.3 Data journals

Institutionalised data publishing in journals is available for scientists for several years. The main way to distribute the data is supplement data to a written publication, which was published in a scientific journal. In this case, an author uses available storage provided by the journal and stores additional data to it. This data is normally also part of the peer review process of the article that it is appended to. A limitation of this way of publication of stored data is given by the fact, that the data should only underline the argumentation of the article (Lawrence et al. [2011]). Therefore, new ways of publication emerged in recent years. Important developments are data journals, which will be discussed in the following.

Data journals are journals, which are specialised on presenting datasets, their retrieval and preparation (AGU Publications Committee [1993]). The focus is set on the dataset itself, which is described in detail by the authors of the data. An example for this new generation of journals is "Earth System Science Data" (ESSD). It was created in 2008 by Hans Pfeiffenberger and David Carlson and is published by

Copernicus. In the publication process, ESSD wants to form an independent publication, which coexists alongside the traditional paper. Therefore, it includes a quality assurance, which consists of an associated discussion paper and a peer review process like the traditional publications (ESSD [2012b]). The only difference is, that it does not reflect the whole traditional scientific process, like it was described in section 2.2.2. It omits the argumentation around a pre-experimental and post-experimental theory and focuses on the experimental design itself. As a consequence, it aims to get cited by a traditional publication to round up the detailed documentation of the whole scientific process (Pfeiffenberger and Carlson [2011]).

An advantage is, that users of the data are able to direct their citation at the dataset itself instead of taking reference to a paper with a different focus. Additionally, a performed quality assurance, which might be described in the data paper, gives the dataset an additional value and therefore for the author of the paper himself/herself. Disadvantages can be found in the fact, that these papers need additional effort by the authors of the data. Like explained before, it is hoped, that invested effort increases their credibility. It is also possible, that the data journal focuses completely on the experimental design, what would not help a data user to correctly access and analyse the data. The format of the short analysis is a free form text just like a traditional paper. This might be of concern for search and retrieve, because it is still discussed, if standardised metadata are better than free form texts for delivering search results in literature databases (Kostoff [2010], Beall [2008], Hemminger et al. [2007]).

A part of the data journal process, which is used by the ESSD, is the data storage at a data repository (Pfeiffenberger and Carlson [2011]). An enhancement of this pure storage functionality of a data repository, the primary data publication, will be explained in the next section.

2.3.4 Primary data publication

In this section, the difference between the storage functionality of a data journal and the primary data publication will be emphasised. Therefore, the traditional publication process, which was shown in figure 2.1 is enhanced. The scheme with this enhancement is presented in figure 2.2. In blue, the elements of the traditional publication process, which were already described in section 2.2.2, are shown. The elements in red show the processes, that were included since the advent of the world wide web. The new processes include the quality assurance step for the traditional publication, which was described in section 2.2.3. Data journals, which were introduced in section 2.3.3, build a new branch of publication. It is similarly structured like the traditional publication. The main difference is, that it does not include a full analysis of the scientific problem. It rather focuses primarily on the experimental design of the experiment. Therefore, the analysis step, which produces the manuscript, is described here as 'short analysis'. The other elements in this branch are equal to the traditional publication, which includes the possibility of quality assurance.

The primary data publication is shown in green. It consists generally of two parts: a section of the data itself and the data on the data, the metadata. Both need different quality assurances, which are particularly described in section 2.4.2. The quality assurance on metadata is marked as "QA/QC", which stands for quality assurance and quality control. The reason will be discussed in section 5.3.1.3. When this quality assurance step is done the data is published. A detailed look at figure 2.2 shows, that when only the green elements are included in the process description of the data publication, one element is missing. The peer review process, which is included in the data journal and the traditional publication process is not defined for primary data publication yet. Therefore, due to the reason of symmetry the element of peer review is included in black. How a peer review process for primary data publication might look like, will be discussed in section 5.3.4.

The difference of the three types of publication cannot only be found in the different forms, but also in the different parts of the publication process that they cover. Besides the data, which are an integral

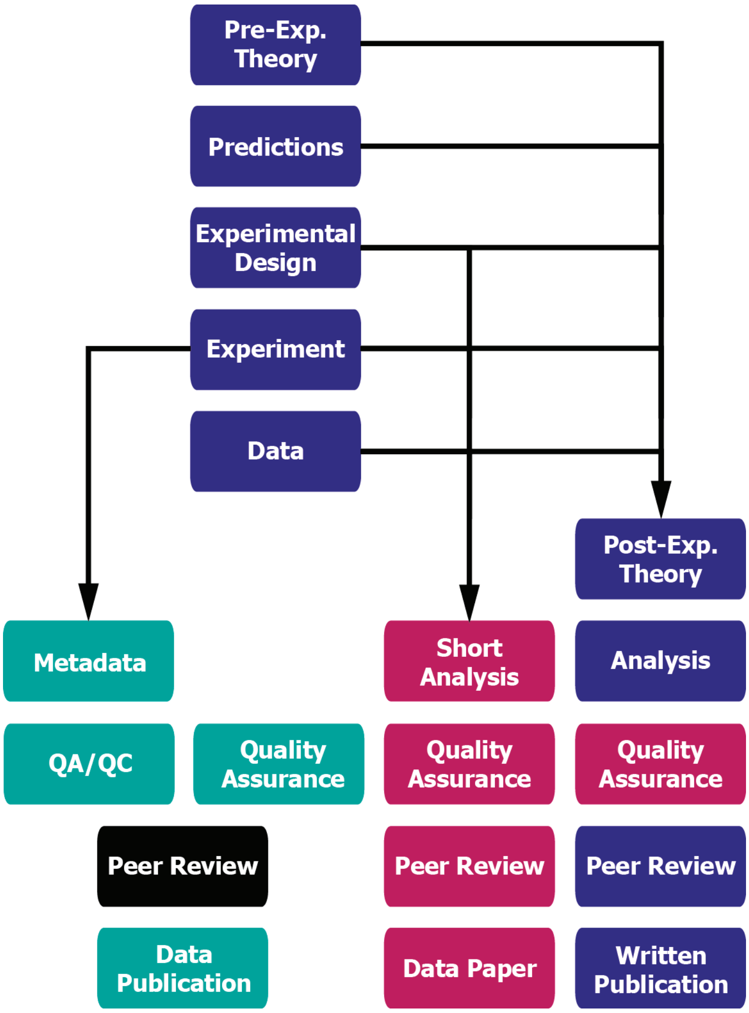


Figure 2.2: An idealised sceintific working scheme. In blue the traditional steps, already shown in figure 2.1. In red the new elements since the advent of the world wide web. Elements of the primary data publication are shown in green. The missing peer review step in black, since this is still under research.

part of this type of publication, the primary data publication covers information on the experiment. This information is stored in the metadata, which will be explained in more detail in section 2.5.5. Data journals cover additionally the experimental design and cite the basic dataset. This dataset can either just be stored at a data repository or be published as a primary data publication before. The whole scientific process is only described in the traditional form of publication. It is possible, that this major form of scientific publication cites a data journal or a primary data publication. Important parts within the process of the three forms of publication are the quality assurance steps, which will be further explained in the next section.

2.4 Quality assurance in the data publication processes

The quality of published entities is crucial for their reusers. Therefore, quality assurance measures have to be defined and implemented into the publication process, which will be demonstrated in this section. It starts with a brief overview in section 2.4.1, on the question, what data quality is. After this, section 2.4.2 explains the properties of the quality assurance processes for the different publication types. Therefore, three different types of quality assurance are defined and explained. Finally, in section 2.4.3, the importance of the quality assurance process is further examined.

2.4.1 What is data quality?

The term data quality is often used, but hard to define. A definition for general databases is given by Wang and Strong [1996], who define data quality "... as data that are fit for use by data consumers". In their study they determine by survey 179 attributes, that contribute to data quality. After the reduction to 118 attributes by refining the survey method, Wang and Strong [1996] group the attributes in four groups: Intrinsic, contextual, representational and accessibility data quality. These cover not only the quality of the data values itself, even when this is mostly associated with data quality (Fox et al. [1994]). Under this point of view data quality also accounts for attributes, which cover the whole generating and archiving process of the data.

In connection to scientific data generation this means, that the data quality has to consider also the "instrumentation, observing practices, data handling and processing procedures, archiving and dissemination" (Guttman and Quayle [1990]). Therefore, quality assurance in data publication has to be set on a broad basis to achieve a result, which leads to an acceptable standard for the data reusers. As a consequence, in this thesis, data quality is therefore not only understood as quality of the data values, but additionally as well documented in all the parts, that were mentioned above. This is reflected in the quality assurance, which will be proposed in the next section.

2.4.2 Different types of quality assurance

In section 2.3.4 four different quality assurance elements were mentioned. The types of quality assurance of a data journal and the traditional publication are similar, since both consist of a free form text. Therefore, three different types of quality assurance have to be performed, if all three branches of publication ought to be passed with an experiment.

In the following, the three types will be characterised. The different elements are shown in the figures 2.3 to 2.5. All three types are divided into three different phases that are represented by colours: collecting the information in green, controlling the information in blue and documenting the quality assurance in red.

The first type of quality assurance to be discussed will be the one of the traditional publication and the

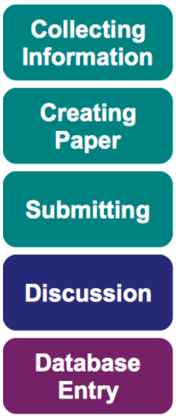


Figure 2.3: Workflow of a proposed scientific quality assurance on methods. Green elements are part of the collecting, blue of the controlling and red of the documenting part of this quality assurance.

data journal. They can both be characterised as free form texts, which follows a recommended form. Still, it has no generally standardised and fixed form. For simplicity, the practical elements of analysis in the traditional publication and short analysis of the data journal are handled as parts of the quality assurance as well. An idealised quality assurance process for this is shown in figure 2.3. It begins with the collection of the information and the production of a manuscript for the analysis or short analysis. When this is done, it can be submitted to a journal. These collecting steps can also be seen as a part of the analysis itself. The following controlling step can be, for example, the production of a discussion paper or a pre-publication, as it was explained in section 2.2.3. It is also possible, that this step includes a first technical control of an employee of the journal, which is not part of the proper peer review. After having finished this controlling step, the performed steps are documented by including them to a database. It is also possible to make this database entry publicly available, for example on the website of the journal or in form of additional information of the paper. This is part of the freedom the scientific journal has. Some further remarks on this flexibility are given in section 5.3.1.2.

As a second form the quality assurance of the metadata, as part of a primary data publication is shown in figure 2.4. Metadata can be characterised as mixed type, highly standardised information. It starts with the collection of the metadata. This can either be done by asking a scientist to report the necessary information or by extracting it from the dataset directly. The latter might be possible, if the data format of the dataset consists of data and metadata. An example for those formats is netCDF (Eaton et al. [2011]). After collecting or extracting this information it might be necessary to complete the metadata. An example for this completion is given in section 2.5.4, when the web-based software Atarrabi is described. After completing the set of metadata, a technical check of the entered information is necessary to make sure, that this information can be included into the database. In a the first step, this is done by an automatic validation. In a second step it is useful, if a human, here called publication agent, rechecks the inclusion to correct obvious errors. These may occur due to misunderstandings of the standardised format, in which metadata have to be compiled. When these corrections are finished, the information is included into a database.

The third type of quality assurance is the quality assurance of the data itself. The corresponding idealised process is shown in figure 2.5. At first data have to be collected. Should a data centre be the storing location, the scientists have to upload the data to it. The data centre stores the information into their database, which includes a first quality measure, because the data have to fit into the database. Now,

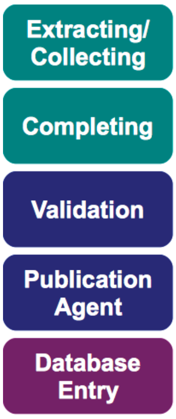


Figure 2.4: Workflow of a proposed scientific quality assurance on metadata. Green elements are part of the collecting, blue of the controlling and red of the documenting part of this quality assurance.

two types of control are applied to the data. The first is the quality assurance, which checks, whether the content of the data is correct. This is called 'Scientific Quality Assurance' (SQA). Since the development of such a SQA will form the main part of this thesis, it is explained in detail in section 2.6. The results of the 'SQA on data' are used to enhance the metadata in a documenting step. When this has been completed, a technical check on the data is performed, which is called 'Technical Quality Assurance' (TQA). When everything is finished, the additional information is appended to the data in the database and the quality assurance on data is completed.

The three types of quality assurances are all necessary, when a dataset shall be quality assured in a trustworthy way. All three show technical and scientific parts and can therefore be divided into a technical quality assurance and a scientific quality assurance. The SQA and the TQA on data were explained above. The SQA on metadata is given by the inclusion of the publication agent, the TQA in the technical validation. In a data journal, the quality assurance is performed on the methods of the experiment. Therefore, the steps are called 'SQA and TQA on methods' in the following. The SQA on methods can be defined as the possible discussion paper, which can be generated within the quality assurance. The TQA on methods can be found in the submitting step to the journal, when the journal checks, whether the free form text fulfils the technical requirements of the journal.

After having separated the quality assurance into the three different types, the question arises, why a scientist and a data centre or journal should invest the efforts to perform a proper quality assurance. This question will be answered in the next section.

2.4.3 Why is quality assurance so important?

Especially in climate science, which strongly builds on data as its foundation, quality assured data are of high importance (Overpeck et al. [2011], Ducré-Robitaille et al. [2003]). Without this data it is impossible to perform an accepted science. In the context of data publishing, quality assurance of data adds value to the dataset (Costello [2009]). This is achieved by offering additional information to a reuser of the data, about whether the dataset fits his/her needs. This allows him/her to optimise his/her scientific workflow.

In environmental sciences the necessity to perform quality assurances on data is recognised for a long time. For example in meteorology this has led to a lot of specialised tests (Wan et al. [2007], Blakeslee and Rumble Jr. [2003]). Nevertheless, testing data and especially measurement data on its quality is

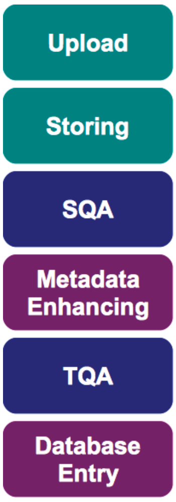


Figure 2.5: Workflow of a proposed scientific quality assurance on data. Green elements are part of the collecting, blue of the controlling and red of the documenting part of this quality assurance.

difficult (Højstrup [1993]). Data publication is a new challenge to this task, since data reusers are not necessarily of the same field that the data originates from (Lide [2007]). For them, it is much harder to decide, whether a dataset has a quality that is sufficient to use it in the aimed application. Therefore, a well performed quality assurance is inevitable.

2.5 Data publication at the WDC for Climate

Primary data publication is located at data centres. Therefore, the data centres' functionality and procedures are explained in this section, with the World Data Center for Climate (WDC-C) used as an example. The system of world data centres is introduced briefly at the beginning in section 2.5.1. Apart from the location, where the data is stored, it is important how the data can be found in the world wide web. One option for this are identifiers, which are presented in section 2.5.2. In the following the publication process which was developed in the project "Publikation Umweltdaten" is shown in section 2.5.3. One of the developments are a web-based software tool for the SQA on metadata, called Atarrabi. An overview on this software is given in section 2.5.4. Some information about the metadata itself and their functionality will be given at the end in section 2.5.5.

2.5.1 World data centers

World data centres were originally founded by the International Council of Scientific Unions (ICSU) in context of the International Geophysical Year (IGY) 1957/1958 (Ruttenberg and Rishbeth [1994]). They exist in several fields of earth sciences and their task is to offer a possibility to collect data from one specified field of research interest. This system was not considered effective anymore in times of a "modern international science" and was therefore transformed to a World Data System (Carlson [2011]). The corresponding data centres store datasets indefinitely and allow scientists around the world to access the data (AGU Publications Committee [1993]).
One of the World Data Centres is the "World Data Center for Climate" (WDC-C) in Hamburg. It was established in 2003 and bases on the technological foundation of the German Climate Computer Centre (Deutsches Klimarechenzentrum, DKRZ). It collects earth system data and sets its focus on data of

climate modelling experiments (Toussaint et al. [2007]). The installed database, called CERA-2 (Climate and Environmental Data Retrieval and Archiving), is used to archive data and to attach additional data to it (Lautenschlager et al. [1998]). In recent years it started to publish primary data. A necessary tool are data handles, which are a topic of the next section.

2.5.2 Referencing data

Archiving data is only the first step of making them available to a broader public. Besides webpages that offer additional information about a dataset, the data have to be reachable in a simple way. A very important instrument, especially for citing the dataset in the literature, are identifiers.
The WDC-C acts as a publication agency for Digital Object Identifier (DOI) (Paskin [2005]). Identifiers are used to make a digital object citable through the assignment of a unique code. By searching this code in a search engine, a web user is able to get directly to a landing page of the dataset (Paskin [2005]). The DOI system was established in 1998 and used by a high number of publishers and data centres (Duerr et al. [2011]). It is not only possible to register DOIs for datasets, but for every available digital document. Hence, they are also used to cite traditional papers and are implemented by aan immense number of journals as a standard identifier. The WDC-C registers the DOI at a non-profit organisation named DataCite, which serves as a global registration agency (Brase [2009], Lawrence et al. [2011]).
With these attributes the DOI is a useful tool to find and cite datasets, and is part of the publication process of the WDC-C. The modelling of this process will be described in the upcoming section.

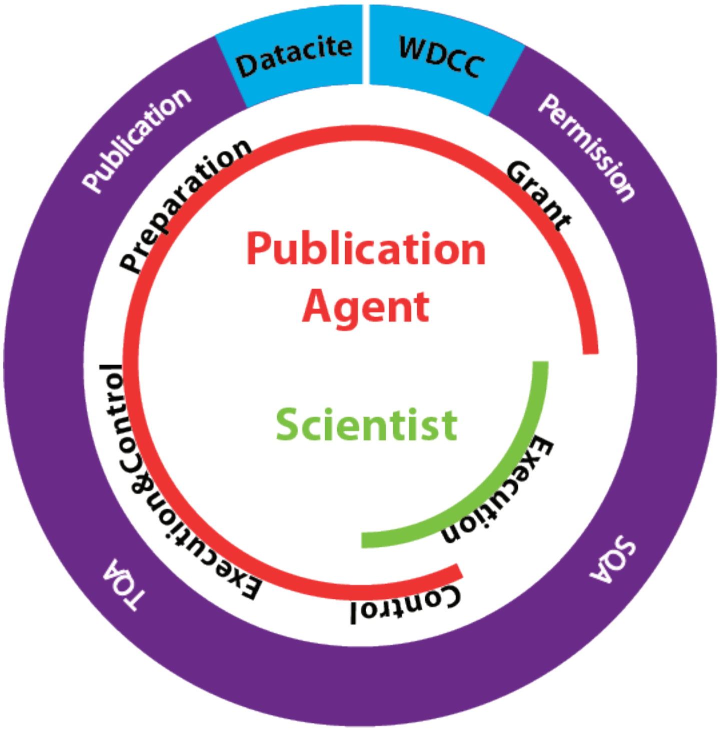


Figure 2.6: Data publication process at the World Data Center of Climate (WDC-C). It starts with the data availability in the WDC-C long term storage and afterwards goes clockwise through the steps, which end with sending the information of the publication to DataCite. It described the role and task of the publication agent (red) and of the publishing scientist (green) by the inner circles.

2.5.3 Modelling the publication process

Within the project "Publikation Umweltdaten" ("Publication of Environmental Data") funded by Deutsche Forschungsgemeinschaft (DFG) the primary data publication process at the WDC-C was enhanced. A basic aim was the inclusion of a quality assurance for data and metadata. To achieve this, the process for data publication was modelled and the tasks of the data centre and the publishing scientists were described in detail (Hense and Quadt [2011]).

The structure of the process is shown in figure 2.6. It shows the process steps in the blue circle, which begins at the top. The starting point of the publication is, that the data is stored at the WDC-C in its long term archive. At the WDC-C, the publication of a specific dataset is in the responsibility of a publication agent. He/She acts as the contact person for the scientist. The working steps, to be performed by the publication agent are indicated by the red inner circle. The steps to be performed by the publishing scientist are shown by the green circle. The specific tasks are added to these inner circles.

The first step is initiated by the publication agent after the WDC-C has decided to publish a specific dataset. He/She starts the workflow by granting the technical permission. This includes an invitation to perform a scientific quality assurance on the data and metadata to the publishing scientist. The scientist starts with the SQA on metadata, for which he/she uses a web-based workflow system called 'Atarrabi'. This will be presented in the next section (2.5.4). In this workflow the scientist is asked to perform the SQA on data and document it within Atarrabi. When the scientist finishes the SQA, the publication agent controls the entered information. The representative of the WDC-C also decides, if the information given in the SQA fulfils the requirements. If not, the publication agent helps the publishing scientist to ensure the needed quality of the SQA.

When the SQA is accepted, the publication agent initiates the technical quality assurance. Should the datasets pass these tests, all information necessary for the data publication is collected. Afterwards, it is send to DataCite, which offers the DOI, in a standardised form. There, the information is used to create the link between the DOI and the storage at the data centre. Then the publication process of the datasets is finished.

Further details of the technical implementation of this model are given in (Quadt et al. [2012]). One integral part, the software Atarrabi, is presented in the following section.

2.5.4 Atarrabi

Atarrabi is a web-based software system. It helps the publishing scientists to perform a scientific quality assurance on metadata. To achieve this, Atarrabi uses a workflow based approach, which divides the different themes of the scientific quality assurance in separate steps:

- General information on the experiment
- Contact information for the authors of the experiment
- Contact information for the leading author
- Contributing institutions
- Relations to other publications
- Spatial and temporal coverage of the experiment
- Information on the used instruments
- Quality of the data

Each step consists of a dedicated view. In each view the publishing scientist is asked to enter the demanded information, which is divided into two kinds: required and optional. The required information is necessary to get the DOI for the publication of the dataset. Additional information, which is marked as optional, ought to motivate the scientist to give a better documentation on his/her dataset.

An important aim of Atarrabi is to assist the publishing author in creating a good documentation. Therefore, several quality measures are implemented into the software. The first explained at this point are lists of values. These lists preselect information for the scientist, if possible. For example contact information for authors are loaded, if available, automatically from the CERA2 database. Therefore, the scientist just has to edit this information, which is not already in the database or out of date. Another quality measure is the visualisation of spatial information on interactive maps. This shows the user of the software another representation of the entered coordinates. It results in a much simpler way for him/her to see possible errors in his/her entered information in this field. The same is achieved by the validation on technical aspects on every entered information.

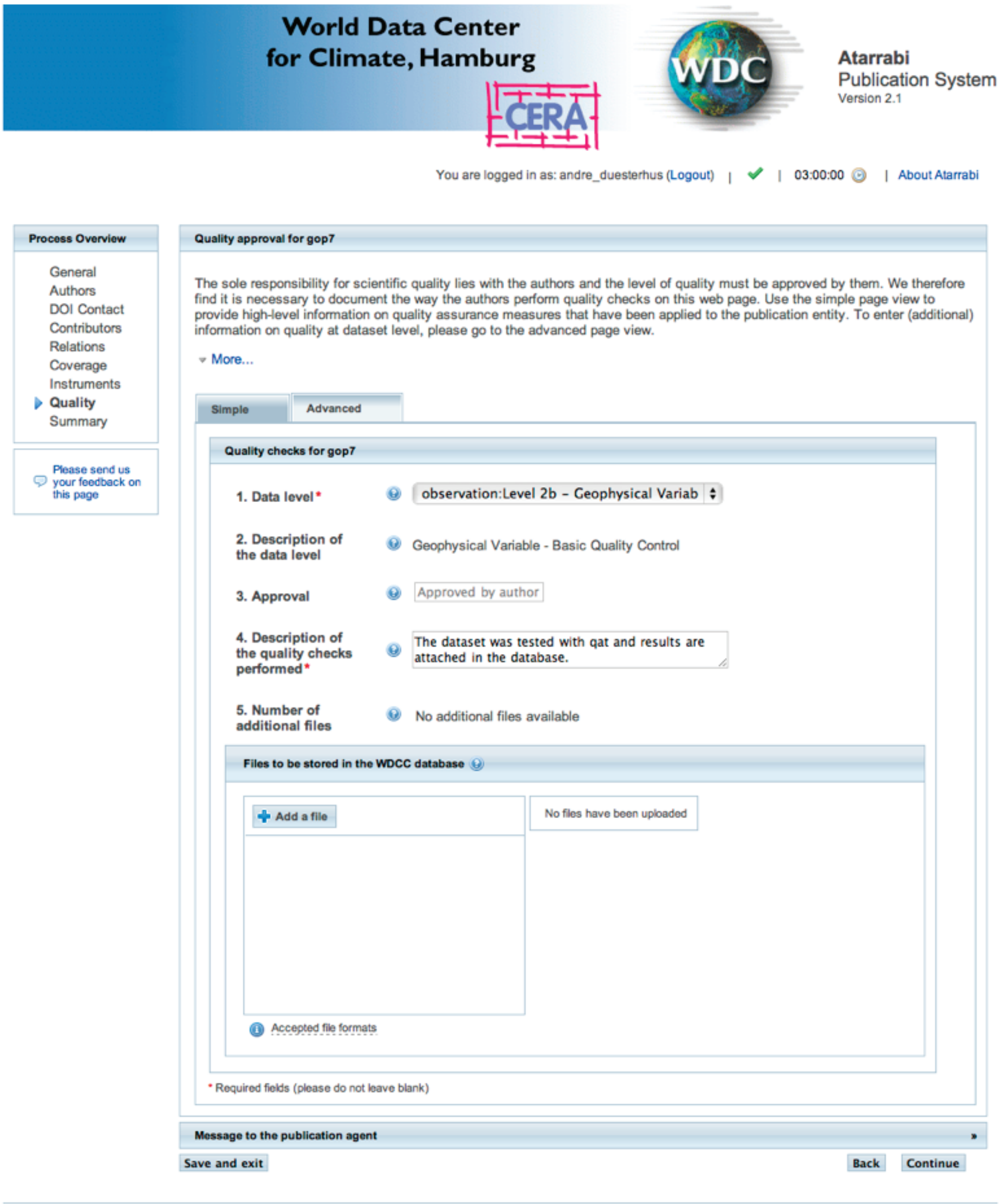
The next quality measure are extensive help texts. These texts include detailed information on what is required in which field. They also inform the user about what happens with his/her information. This should prevent the scientist from misunderstanding the instructions and also motivates him/her to invest more effort in giving a documentation on his/her datasets, which is as good as possible. It also helps to give the scientist a deeper insight into the underlying processes, what increases the transparency of the process (see also section 5.3.2).

A last measure to enhance the quality of the documentation are integrated contact possibilities. These should simplify the communication between the publishing scientist and the publication agent. The latter has an administration panel, which helps him/her to follow the edits by the scientist and intervene if necessary.

Of particular interest for this thesis is the last point of the enumeration above: Quality of the data. This view connects the SQA on metadata with the SQA on data. In figure 2.7 it is shown how this view is designed in the Atarrabi version 2.1. Next to the navigation on the left hand side there is the main box named "Quality approval for gop7". In this case gop7 is the name of the actual project, whose data quality should be documented in this view. It starts with a brief explanation for the publishing scientist, which can be enlarged by clicking the link text "More...". Below this introduction, the user has the opportunity to decide between a simple and an advanced view. The second is designed just like the first, but enables the user to perform the following steps not only for the whole experiment, but also for every individual dataset. In the simple view five steps are shown. The first is the selection of the data level, which is explained in more detail in section 2.6.2. If the user selects a quality level for the experiment, the dedicated description is shown in the second line. In a third point the author gives an approval that he/she is responsible for the quality of the datasets. The fourth line is a textfield, in which the publishing scientist has to enter some comments on the performed quality checks. In a last point the user can upload some files to underline the comments, that he/she had given before. For that, standardised formats like reports in the Portable Document Format (PDF) can be used. It is also a possible interface for uploading results of the SQA on data described in section 2.6.3. By pressing the button 'continue' the user finishes this view.

Afterwards, the publishing scientist is able to examine a summary of all given information in the workflow process. By accepting this summary the publication agent is informed and controls all entered information. When he/she accepts, the SQA for data and metadata ends. Atarrabi itself is a useful tool, because the input of this information has to be done by submitting an Extendable Markup Language (XML) file before its introduction. With the included quality measures it helps to get a better result of the SQA on metadata.

A connection point between the SQA on data and metadata is described from a more theoretical point of view in section 2.6. In the next section follow some remarks on the metadata, which are stored in this process.



You are logged in as: andre_duesterhus (Logout) | ✓ | 03:00:00 |

Quality approval for gop7

The sole responsibility for scientific quality lies with the authors and the level of quality must be approved by them. We therefore find it is necessary to document the way the authors perform quality checks on this web page. Use the simple page view to provide high-level information on quality assurance measures that have been applied to the publication entity. To enter (additional) information on quality at dataset level, please go to the advanced page view.

Quality checks for gop7

1. Data level* observation:Level 2b – Geophysical Variab

2. Description of the data level Geophysical Variable - Basic Quality Control

3. Approval Approved by author

4. Description of the quality checks performed* The dataset was tested with qat and results are attached in the database.

5. Number of additional files No additional files available

Files to be stored in the WDCC database

No files have been uploaded

* Required fields (please do not leave blank)

Message to the publication agent

Figure 2.7: Screenshot of the data quality view of the software Atarrabi (version 2.1).

2.5.5 Metadata

Metadata are an important addition to the primary data. They are defined as "data about data" (Peterson et al. [1998]) or "information which makes data useful" (Bretherton and Singley [1994]). The information, which should be stored in the metadata depends on the reuser of the data. Bretherton and Singley [1994] described the different meanings of metadata for scientists from different fields. For example, a computer scientist understands under the term metadata "physical level information", like for example file names. Nevertheless, with the term metadata physicists would associate information, which is necessary to understand the content of the primary data. This could be for example the used instruments of the experiment, in which the dataset was generated.

That well compiled metadata are important, was stated for example by the AGU Council [2009]: "Because datasets are often later used for purposes other than those for which they were collected, accurate, complete, and, when possible, standardized metadata are as important as the data themselves."

In data publishing, metadata are especially important even when it is hard to find the balance between too much needed effort and usable information for a data reuser (Lawrence et al. [2011]). As a basic convention for metadata DublinCore can be used (Lawrence et al. [2011], Weibel [1997]). This defines a collection of required and optional entries for a set of metadata. For the publication of environmental data a lot of metadata can be generated and stored. An example is the author of the data. The first question is: Who has to be named as an author of the data? Parsons et al. [2010] defined data author with: "Authors are those who put the intellectual effort into collecting and preparing the data". For the people who fulfil this requirement, a name for registering the DOI is required. Nevertheless, it might be of interest for a data centre to ask for more details to a person, like the institution these people belong to, a contact address, phone numbers etc. As a consequence, it is hard to define a limit on which information is necessary and should be asked from the publishing scientists.

Within the project of "Publikation Umweltdaten" a basic set was defined, which covers the themes shown as steps within Atarrabi in section 2.5.4. This basic set was enriched by additional optional information. To prevent a user from filling out entries, which can be generated by already known information of the data centre, he/she is assisted by the software. When he/she enters a name, the database of the data centre will check, if it is an already known person. If so, additional information will be filled in automatically, which the user just has to control and accept. With such measures the time, which is needed by the user and data centre staff to produce consistent and correct metadata, is clearly reduced.

2.6 Scientific quality assurance of primary data

After having explained the environment and necessity of a scientific quality assurance on data this section will show, how such a process can look like. It starts in section 2.6.1 with a definition on what an SQA on data is and which recommendation follows for the processes involved. Afterwards, section 2.6.2 shows the importance and possibilities of a well performed documentation of an SQA on data. To assist a user with the performance and documentation of the process, a proof of concept implementation of a software is shown in section 2.6.3. This software bases on quality tests on data. To be able to develop these, the basis is laid in the following. It starts with reasons for errors in datasets in meteorology in section 2.6.4. These have consequences for the resulting datasets, what is explained in section 2.6.5. Based on that, the necessity of generalised tests is briefly discussed in section 2.6.6. Before tests can be developed, it is necessary to think about the evaluation of the results. Since a lot of tests lead to a lot of results, which have to be effectively evaluated, a quality evaluation model is explained in the last section 2.6.7. From this model, some prerequisites follow for the tests in the upcoming chapter 3, which then describes and develops some general quality checks.

2.6.1 What is primary data scientific quality assurance?

The starting point in this section is the question on what a primary data scientific quality assurance is. It will be explained, which tasks it should handle and which basic problems it should solve.

The basic characteristic of the SQA on data should be test-based. This means, that if a scientist performs the SQA on data he/she ought use, for example statistical, tests to check the data quality. This is necessary, because it simplifies the reconstruction and documentation of the performed process. These tests provide results, which then have to be interpreted by the user. If tests provide indications for suspicious data within the dataset, the publishing scientist should comment on this and try to explain the reasons for their existence. The main aim is to give an indication to the data reuser if he/she can use the dataset for his/her purpose. It should also give him/her a starting point for the search of errors, if the data behave strange in the reusers own analysis.

To fulfil these aims, it is necessary that within the quality assurance process the data are not changed themselves. This is different to other systems that follow more the idea of a quality control. In these systems a change to the data itself is a basic aim, what gives a more accurate representation of the measured truth. This is discussed in more detail in section 5.3.1.3. Keeping the original datasets is important, what was also stated by You and Hubbard [2006]. This statement does not mean, that corrected datasets can not be published at all. It just states, that the process to perform the test of quality within the publication process should not apply such corrections. If a publishing scientist wants to publish corrected datasets he/she has to deliver the data centre these corrected datasets as new publication entities, which then have to go through the quality assurance step again.[2]

Another problem, which has to be solved by the quality assurance system is the handling of immense amounts of data. A usual approach would be to only look at a subsample of the dataset and then extrapolate the results. Since it is the aim to generate confidence of the quality statement to the whole dataset, all data have to be checked. Quality checks themselves only control a subset of a dataset, sometimes handle every datum individually (Hubbard et al. [2005]). This means, that finding a quality statement on the whole dataset can be a complicated task. As a consequence, there is a need of mechanisms to allow the publishing scientist to perform this type of SQA as convenient as possible. A way to achieve this is the automatisation of quality checks (Gronell and Wijffels [2008], Guttman and Quayle [1990]), which will be thematised in section 2.6.7.

The presented basic structure of a scientific quality assurance on data allows to generate a transparent and comprehensible information for data reusers. This allows to help scientists to use and to interpret the data correctly.

2.6.2 Documenting quality assurance on data

For the scientific quality assurance of data three basic things are necessary to implement: roles for documentation, quality checks and automatisation algorithms. All this will be explained in the upcoming sections.

The first component is the documentation of the quality assurance, whose importance is mentioned for example by Overpeck et al. [2011]. In their opinion, the used quality control procedures should be communicated like changes of the instruments or "spatial or temporal sampling uncertainties". With the detailed documentation of the quality assurance, it is also possible to evaluate the used methods afterwards like it is for example done by Durre et al. [2008].

The concept of documentation is shown in figure 2.8. It starts on the top with the data, which is used in the following to perform the quality tests on it. In a next step follows the documentation. This is also test-wise and created with the aim to make it possible to reproduce every test separately. Therefore, several components have to be stored for every performed test. The first is a name of the test, which can be used for simple identification. A detailed description about what the test is actually doing and testing for, will help interested data reusers to understand the behaviour and the aim of the test. Another

[2] The corrected dataset can also be published alongside the uncorrected data as an additional variable within a new dataset. Nevertheless, this dataset have to pass again the quality assurance as a new publication entity.

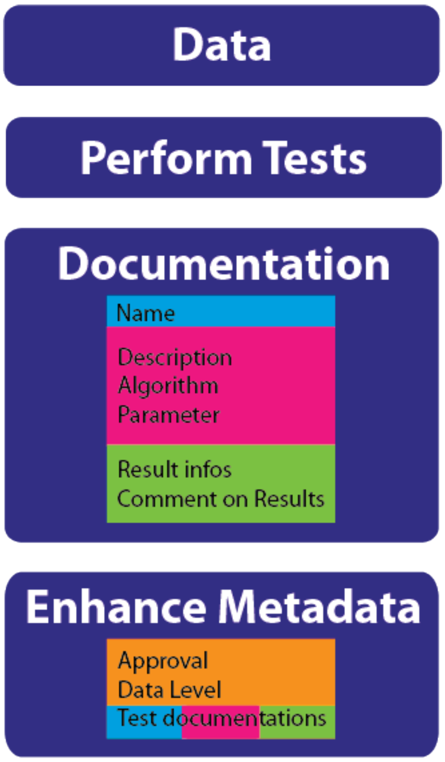


Figure 2.8: Workflow of the scientific quality assurance on data.

important information to store is the used algorithm of the test. This could be the code or a reference to a publication underlying the test.

With this information, it should be possible to exactly reproduce the performed check, if the data is available. The reason for the importance of the algorithm lies in the problem, that small deviations in the construction of the test might lead to different results. As it is shown in section 2.6.6, the tests are mostly constructed in a way, that they are controlled by only a low number of parameters.

Since these parameters are also necessary for a possible repetition of the exact check, they have to be stored as well. The next component is an optional documentation of the results of the check. If the test can be reproduced by the information given for the algorithm and parameters, it is not necessary to store the results. Nevertheless, if for example the computational time to perform a check is high, it might be recommendable to document it anyhow. Examples for types of results are given in section 2.6.3. More important to document is a comment on the results. With this component, the intention of the quality check becomes clear and a data reuser learns, how he/she should interpret the results. This is necessary, because the occurrence of a quality check with bad results does not mean that the dataset itself is of bad quality. If parameters are chosen in a way, that it is hardly possible for a dataset to pass it, the data reuser has to be informed.

All these documentation steps are then used to enhance the metadata, together with additional information. An example for this additional information is the data level of a dataset. Data levels were used for example in the First GARP Global Experiment (FGGE) for observational datasets (U.S. FGGE Project Office Staff [1978]) and are of common use for satellite products (WMO [2012]). They indicate the level of processing and the extent of the performed quality assurance on the data. For observational data at the WDC-C the definitions of the Global Ocean Data Assimilation Experiment (GODAE) are used (GODAE [2007]). Those levels range from 0 for raw instrument data to level 4 for highly processed data.

Added are an a, b or c for a low, medium or high action of the quality control, that was performed on the datasets. Therefore, a dataset with level 2b consists of geophysical variables, which was subject to some action of quality control.

Another additional information, that could enhance the metadata is an approval. This approval consists of a small text, which states who approved the quality assurance results. In the simplest form this is "approved by author", which indicates, that the publishing scientist is responsible. In the future it could be "peer reviewed" as well, if the datasets undergo a peer review process of data.

All the above described information is added to the metadata. A possible way is an interface within the SQA on metadata just like it is done in Atarrabi (section 2.5.4).

This information is necessary to replicate the performed tests and to estimate their influence on the work for a data reuser. Nevertheless, it is a lot of effort for a scientist to perform such a detailed documentation. Therefore, it is necessary to assist him, for example with software solutions. An example for such a software is shown in the next section. It does not only perform tests, but also gives information for a documentation of the performed quality assurance steps.

2.6.3 Implementation of the quality assurance toolkit (qat)

As it was shown in section 2.5.4 in the project "Publikation Umweltdaten" with Atarrabi a software was developed for the SQA on metadata. Since for data publication an SQA on data has to be performed as well, there is also a need to develop a software for this. It will be presented in this section.

A prerequisite by the staff of the data centre, the German Climate Computer Center (DKRZ), was that they do not have to perform necessary calculations within the SQA on data on their own resources. As a consequence, the concept of this type of SQA includes, that the author has to perform the SQA on data on his/her own resources. This is achieved by the provision of an extension package for the statistical programming language R (R Development Core Team [2011]). In this section, the structure of this package will be explained. Other topics are the precautions for the connection between the package and Atarrabi.

The package is named Quality Assurance Toolkit (qat) (Düsterhus [2011]). It is constructed as a proof of concept implementation and focuses on a high flexibility to make it a usable tool for data analysis. It also gives the possibility to include new developed methods for data publication in a simple way. The basic structure is shown in figure 2.9. It consists of twelve modules in four vertical levels. Each module contains several functions to provide the needed functionality. The structure allows the processing of a workflow of tests, what helps the user to perform a quality assurance on a larger number of datasets effectively.

The first module on the left hand side is the reading module. Its function is defined in two ways. At first, it should read the datasets to be controlled. As an example file format, netCDF files are used. They allow to store several variables into one file. The metadata of the file contain information to uniquely identify each variable in the file, what can be used to distinguish between the information in the storing procedures in the writing module. The second functionality is the reading of the workflow setup. For this an XML scheme is used, which contains information, like parameters, for the tests, that should be performed. For the upcoming steps the dataset, which ought to be controlled, has to be divided into logical entities. This is in its simplest form a one or more dimensional time series. In the following, those time series will be called measurement vector (even though this could also be a field). The information, which is given to the next process steps by the reading module is a measurement vector and a dedicated workflow description.

The next part of the package is controlled by a moderator module. This module calls the processes, which perform the tests (analyse), produce plots of the results (plot) and prepare the results for storing (save).

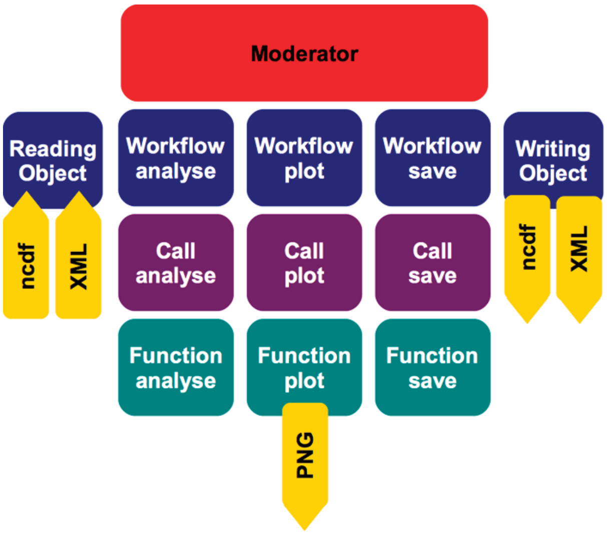


Figure 2.9: Structure of the software package quality assurance toolkit (qat).

The three vertical levels below the moderator are equally structured for all three process types. The first level is the workflow level. This level uses the given workflow, defined by the XML scheme. Its task is to call a different function for each test at the second level. These functions are named calling functions. The main task of the second level is to call, if necessary, different functions for the third level. Reasons for using different functions on the lowest level can be several types of parameters, that request different functions. This concept can also be used to call different functions for measurement vectors with different dimensionality.

A typical processing of a quality check starts with the analysis. The workflow is processed and for every test the calling- and function level is used to get a result for every workflow step. Additionally, the XML of the workflow is enhanced with information on name, description and algorithm. The results of the tests are collected separately.

After having finished all tests, the results are used to start the plot process. Here, the two lower levels, which produce the plots, are called as well. In this proof of concept version, these plots are stored in the Portable Network Graphics (PNG) format.

In a last step, the results and the enhanced XML of the workflow are used to prepare the storing of all results. They produce an output, which merges information for the documentation and the results of the performed checks.

The last module is the writing module. It writes the information generated in the main process into files. The first is the enhanced XML file of the workflow, which now includes information on the name, description and algorithm of the performed tests. It also processes the information generated within the 'save' process to produce an output, which consists of the documentation like the XML, but also of the results of the tests. This is stored in the netCDF format, for which an adapted definition was created.

With this structure, it is possible to perform a quality check for one measurement vector. If several measurement vectors are part of one dataset, the process has to be repeated for each of them. A useful identifier to separate the results of the different measurement vectors is, apart of the filename, the identification, which was already mentioned above. In the case, that a numerical numeration exists within the file, a number of the variable can be used as well.

The produced files, the enhanced workflow in XML, the documented test results in netCDF and the produced plots in png can then be used within the interface in the SQA on metadata. In case of the software Atarrabi it is the file upload in the quality view (section 2.5.4).

The structure shown here can be expanded, like it will be proposed in section 5.2.1. It is also formulated in general, so that it can be used as a basic procedure description to implement such a software in other environments. This is proposed in section 5.3.3.2. In the following section, the basis for the quality checks is determined by giving indications, for which sources of errors have to be covered.

2.6.4 Reasons for errors in data

The environment of the quality assurance is set by the software explained for the SQA on metadata in section 2.5.4 and for the SQA on Data in section 2.6.3. The next step is to find quality checks, that can be used in the analysis process of qat. Therein, it is important to be aware of the typical sources for the errors, that should be detected within the dataset.

The possible reasons for errors obviously depend on the type of experiment, the used instruments and the processing of the data. These sources have to be analysed in detail for every performed experiment and can therefore not be generalised. Nevertheless, some sources are common and well described in the literature.

Durre et al. [2010] described the sources of errors for the Global Historical Climatology Network (GHCN) with "...variety of measurement, recording, digitisation, transmission, and processing problems". Kunkel et al. [2005] used a more hierarchical structuring of the sources of errors in the Cooperative Observer Network (COOP), which consists of three classes. The first, described by "observer errors", cover problems with the functionality of the instruments, the reading and the documenting of the observed parameters. With the second class, "station discontinuity", the problems in datasets are introduced by "...changes in instrumentation/shields, observing practices, changes in station location, and exposure." Those errors are common in climatological datasets. The third class, "digitization errors", discusses identifier problems of stations or measured parameters and "...keying errors in individual values". In homogenisation of surface wind observation by Jiménez et al. [2010], problems of recalibrating instruments, change of sensors, or spatial and temporal inhomogeneities of the observation are mentioned as problems.

All these sources deliver anomalies in the dataset, which have to be detected statistically. These statistical types of anomalies in data are explained in the upcoming section.

2.6.5 Types of errors in data

As a consequence of the possible errors discussed in the last section, the focus of this section lays on the consequences for the datasets, that have to be analysed. Since in meteorological and climatological sciences most datasets consist of time series (see also section 2.1.2), this section will focus on time series analysis.

Classification of the different types of errors is given by Gandin [1988]. He divides them into three basic categories: "Random, systematic, and rough errors". Random errors are the inherent uncertainties of measurements, since measurements are just an approximation of the real physical parameters. They are statistically characterised by their distribution, with a mean at zero.

Errors with a non-zero mean are classified as systematic errors. Just like the random errors the systematic errors are persistent in the time series throughout the measurement, when the procedure of the experiment is not changed. The last type are the rough errors, sometimes also called gross or large errors (Zahumensky [2007]), which are non-persistent in time. A prominent candidate for gross errors are outliers. A fourth type described by Gandin [1988] are micrometeorological, also called representativeness errors (Zahumensky [2007]). They are introduced into the results by small perturbations of the measurement environment and are therefore hardly distinguishable from random errors. As a consequence, Gandin [1988] defines random errors as a combination of the original random error of the observation itself and the micrometeorological errors.

In quality control, the target of the used procedures is to detect and/or reduce systematic and rough errors. Typical rough errors are outliers and missing data. Systematic errors are for example changes in the measured statistical moments of the data, like mean or variance, or introduced trends. Also the possibility of different rounding of the data can be an indicator for an error within the dataset. When these errors occur within a dataset under control, it will in the following be described as an inconsistency. Searching for these systematic errors, when their inclusion occurs outside the time series under control, is only possible with more information. This is for example done in homogenisation, where a lot of methods use reference stations to detect these errors (Easterling and Peterson [1995]).

Testing for the above described errors can either be done with very specified tests on the datasets or by general quality checks. The use of the latter will be motivated in the next section.

2.6.6 The need for general quality tests

Testing meteorological time series for errors is a common task for meteorologists. For that, a lot of quality checks exist in the field of meteorology and climatology. At first glance, most of them seem to be very specialised in terms of their field of application. An example is the quality control of radio sondes, which are a common application for quality control tests (Wan et al. [2007]). Gandin [1988] used checks, that base on the hydrostatic approximation. Information of this approximation is used to define limits for temperatures dependent on the height of the measurement and the temperature at the base level. Looking at this in detail shows, that it is basically a simple limit check with dynamical limits (see also section 3.1.1). The only additional effort that is necessary, is the transformation of the data to the limits to be checked.

This example illustrates, that most of the quality checks can be led back to simpler, more general quality tests. The reduction of checks to such general quality tests is of great importance. First of all, they simplify the understanding of quality tests for data reusers. In a working environment like data publication this is crucial. Only, if it is possible for a user of the data to understand the used tests, he/she will be able to estimate, if the data is usable for him/her. Furthermore, it helps a user to trust the quality estimation and as a consequence the data as well.

Another advantage can be found in the documentation of quality checks. In this phase of a quality assurance, it is sometimes very complicate to transfer the information on what has been done to check the quality to a simple understandable form. By the modification of general tests it becomes possible for the controller of the quality to make this documentation in a standardised form.

A third advantage is the possibility of a simpler reprogramming of the quality tests. This can be done for example by the data reuser to reperform the quality check with a slightly different set of the test driving parameters. It also helps to use a standardised code in data centres, which works with a good performance on the existent computer architecture (see also section 5.3.3).

All this only works, if quality tests are standardised. Doing this standardisation for so many variables in a lot of different environments separately, is a nearly impossible task. Therefore, it is required to collect, to develop and to use general tests, which can be modified by changing the test driving parameters.

Additionally to the tests, there have to be automatisation algorithms, that help to evaluate the outcome of the checks. Such a procedure is proposed in the next section.

2.6.7 Primary data quality evaluation

In this section a procedure is described, which allows to evaluate the quality in the case of the use of several quality checks. That such procedures are necessary, is described for example by Gandin [1988]. In his "Complex Quality Control (CQC)" he uses several different types of quality tests, but as a "main principle" he requires that "no decision is to be made until all available QC methods have been applied to the data under consideration".

The quality evaluation procedure discussed here bases on two assumptions. First, that the tests used in this framework are able to deliver a probability that a dataset passes the test. How this is achieved depends on the type of the tests. Possibilities for fulfilling these requirements can be found in section 5.2.2. The second assumption is the existence of an expert, who is able to evaluate the performance of the dataset in a test and who is able to make a connection between the results of a test and a reference for the quality of the data. Both assumptions are not simple to achieve. Especially the second assumption needs further discussion, that can be found in section 5.2.2.

Should both approaches be valid, the procedure can be set up to calculate a quantity Q, which will be later used to identify the quality of the dataset. The quantity Q depends on the measured observation O, which represents the values of the dataset. It also depends on the real value of the measured quantity, which is the unknown truth T. An occurring problem is, that observations only approximate the truth due to limitations of the used instruments (Gandin [1988]). As a consequence, it is necessary to introduce a measurement operator M_O, which transforms the truth T to an observable value. To get information on the quantity Q, tests should be used. Especially generalised tests depend on their driving parameters θ. They are used to minimise the effort to reprogramm quality tests in order to adjust them to new fields of applications. Examples are the minimum and maximum limits in a test on limits, which will be explained in the description of the LIM-test in section 3.1.1. Here, it is important, that the test is completely defined by the set of parameters θ. With the use of the marginalisation theorem the following equation can be calculated:

$$p\left(Q|O,M_O(T)\right) = \int\limits_{\theta} p\left(Q|\theta,O,M_O(T)\right) p\left(\theta|O,M_O(T)\right) d\theta. \tag{2.1}$$

The other information to be used is the a priori knowledge of experts about the data, given as the property I. To introduce I into the second term of the equation above the marginalisation theorem is used again:

$$p\left(Q|O,M_O(T)\right) = \int\limits_{\theta} p\left(Q|\theta,O,M_O(T)\right) \int\limits_{I} p\left(\theta|I,O,M_O(T)\right) p\left(I|O,M_O(T)\right) dId\theta. \tag{2.2}$$

Both the parameter sets θ and the information of the experts I are discrete and therefore a discretisation is appropriate:

$$p\left(Q|O,M_O(T)\right) = \sum_i p\left(Q|\theta_i,O,M_O(T)\right) \sum_j p\left(\theta_i|I_j,O,M_O(T)\right) p\left(I_j|O,M_O(T)\right). \tag{2.3}$$

As a consequence, the final equation consists of four probabilities, which are interpreted in the following. On the left hand side the probability for the quantity Q is given, if the observation and the modified truth are known. If Q is defined as a good quality of the observations, then this term can be identified as the probability for a good quality of the dataset. On the right hand side the first term is the probability $p\left(Q|\theta_i,O,M_O(T)\right)$. The latter can be identified as a test, since it gives a probability for a good quality

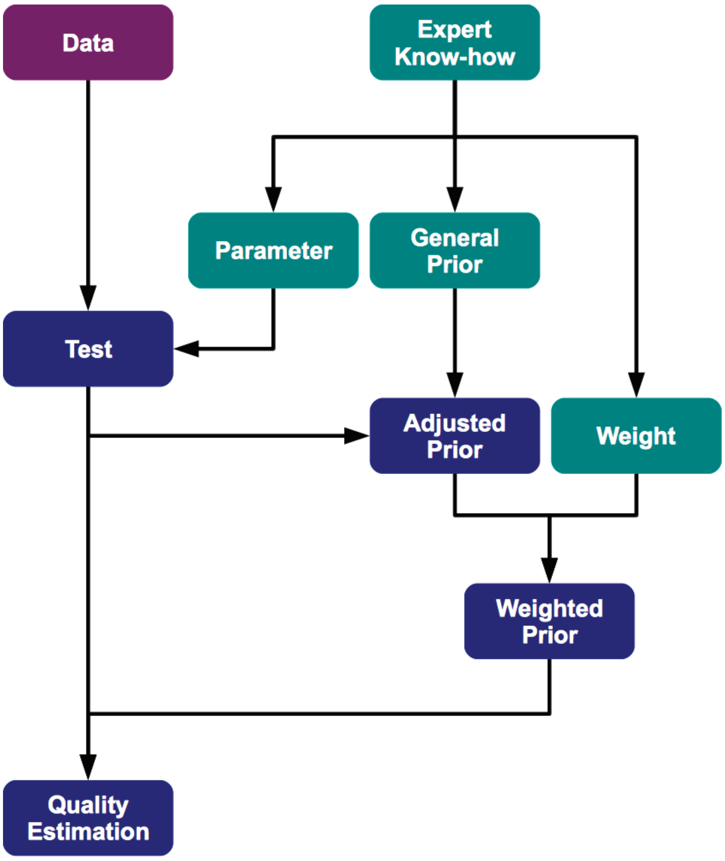


Figure 2.10: Workflow diagram of the quality evaluation process. Green are the elements, which are only influenced by the expert knowledge, red the elements only influenced by the data and blue the elements, which are influenced by both.

of the dataset, dependent on the parameters of the test θ, the observations O and the modified truth $M_O(T)$. The other two terms are the prior information. The second term on the right hand side is the prior of the parameter set in dependence of the knowledge of the expert, among others. It is a weighting of the used set of parameters. The third term $p\left(I_j|O, M_O(T)\right)$ is the prior information of the observation by the expert.

The advantages and disadvantages of this statistical model will be discussed in section 5.2.2. In a next step, a technical implementation of this framework will be demonstrated. To illustrate the process, a structural diagram is shown in figure 2.10. It shows an acyclic directed graph (Pearl [2000], p. 12), which is valid for a set of parameters as well as information by an expert. By implementing the use of more parameter sets or several experts, it has to be modified to a cyclic graph. Nevertheless, in this case the process steps are the same. The additional measures, which have to be performed in such a case, are also explained in the following.

The basis of the process shown in figure 2.10 are the data, indicated in red on the upper left and the expert know-how in green on the right. The process starts with the expert, who has to deliver three things. As a basis for the tests he/she has to define the parameter sets. For every parameter, two additional pieces of information have to be given. First the general prior. It consists of a function, which delivers a prior for every possible outcome of the test with the parameter set. By definition, the outcome and the prior of the test should be ranged each between 0 and 1. Secondly, a weighting factor for the parameter set should be given. In the case of more than one parameter set, the weighting factor can be used to give priority to a certain parameter set. If more than one expert takes part in this process, it can also be used to weight the influence of each scientist on the quality estimate according their expected knowledge about the dataset. To obtain appropriate results, the weightings of all used parameter sets within the quality evaluation should sum up to 1.

With these three pieces of information the testing procedure starts. First, the data is tested by the quality check, what delivers a probability for succeeding the test. Afterwards, this probability is used to get the dedicated prior from the general prior. This is, by definition, a value between 0 and 1 and named adjusted prior. By multiplying this adjusted prior with the weighting, the weighting prior is formed. In a last step the weighting prior is multiplied with the probability of the test, what delivers the quality estimate of the data for a test with the given parameter sets.

For more than one set of parameters all quality estimates of all parameters are summed up to get the resulting quality estimation for the dataset as it is shown in equation 2.3. This is by definition of all used components of this procedure a value in the range from 0 to 1. According to the definition of Q, a good quality is achieved with high values of the quality estimation.

To get a full quality estimation, in theory all possible tests with all possible parameters and all possible knowledge of the experts have to be performed with a dataset. Since this is impossible, the quality estimation is just an approximation of the real quality estimation of the dataset. It is also highly subjective, especially due to the fact, that the expert does not only define the parameters, but also the prior and the weighting. Therefore, this process has to be comprehensively discussed, what will be take place in section 5.2.2.

The part that still has to be defined are the quality tests. This follows in the upcoming chapter 3.

3 Methods

In the last chapter, the general framework of a scientific working process within data publication was presented and the necessity to develop quality tests for general data was emphasised. Those tests should be driven by parameters and be applicable to a wide range of datasets. Since most datasets in meteorological and climatological science consist of one- or multidimensional time series, this chapter will focus on time series analysis.

The first tests will deal with the ones described by Meek and Hatfield [1994] in section 3.1. These are simple tests, which are designed for the application to one dimensional time series. The basic tests will be enhanced with some modifications to make them a general tool for quality tests in even more fields of applications. The second test group are tests working on statistical moments and parameters, which are presented in section 3.2. With these tests it is possible to look at the developments of their underlying distributions. Both groups of tests are introduced briefly in this chapter, since they are simple examples for general testing methods of data. They are also used in some applications in chapter 4.

A more detailed analysis will be performed on the next two types of tests, which are more complex. A special test, which works with histograms as the estimation of the distribution of the dataset and their development, will be introduced in section 3.3. There, several sensitivity test will show how the test might be applicable to general datasets. In section 3.4, a probabilistic change point detection system will be presented, which was developed by Dose and Menzel [2004]. Here, sensitivity tests are used and comparisons to other change point detection systems are performed. In the last section 3.5, the histogram test and the change point detection will be combined. The results of this combination will also be used in some sensitivity tests. The presented tests will be utilised for some applications in chapter 4.

3.1 Methods by Meek & Hatfield

The prototypes of general tests were put into a framework by Meek and Hatfield [1994]. They formulate three simple test types, which are used in a lot of applications before and after their publication. These test types are a test on limits (LIM), a test on the rate of change (ROC), and a test on no changing values of the data (NOC). All three basic types of tests are in common use at data centres to search for errors in datasets (Hubbard et al. [2005], Durre et al. [2010], Reek et al. [1992]). They are also part of the recommendations by the "World Meteorological Organization" (WMO) for the quality control of automatic weather stations (Zahumensky [2007]). This section presents the three types and shows, which enhancements can be used to make this general test applicable for a wide range of applications.

3.1.1 LIM-test

The test on limits of the data checks every data point on whether it exceeds a predefined range of values. The basic procedure described by Meek and Hatfield [1994] used fixed limits for the whole dataset. This is named as "LIM static" in this thesis and can be phrased as:

$$f_i = (x_i > a_{max}) \vee (x_i < a_{min}). \tag{3.1}$$

In this formulation f_i is a flag vector and $\vee$ the logical disjunction. If the data value x_i is greater in value than the maximum limit a_{max} or lower than the minimum limit a_{min}, the value is flagged in the flag vector (true and false or 1 and 0). If one of the two limits is not set by the scientist, it will be interpreted as ∞ or $-\infty$, respectively. After checking the whole vector, the flag vector f_i can be evaluated.
In the basic configuration of this test the limits a_{min} and a_{max} are determined by the performing scientist. In a first enhancement this can directly be calculated from the dataset. This "LIM sigma" test uses the standard deviation to define the limits and was used by Hubbard et al. [2005]. To generalise this basic idea it can be formulated with a factor of the standard deviation s, that an outlier is maximally allowed to deviate from the mean. This leads to the following form:

$$f_i = (x_i > \mu_x + s\sigma_x) \vee (x_i < \mu_x - s\sigma_x). \tag{3.2}$$

Here, again the f_i is the resulting flagvector. μ_x and σ_x are the mean and standard deviations of the whole dataset X, respectively.
A third type presented here is called "LIM dynamic", which does not use a fixed limit for the whole dataset, but a dynamic. This can be formulated as

$$f_i = (x_i > a_{max,i}) \vee (x_i < a_{min,i}). \tag{3.3}$$

In this case, for every element of X, both a maximum and a minimum limit, are defined separately. With this definition, the test can account for diurnal or annular cycles within the dataset.
Technically not a fourth type, but also a possible enhancement for testing, is the modification of the data vector X before the test. This can be introduced to all tests by using a modification function Λ and replace all x_i by $\Lambda(x_i)$. An example was already described in section 2.6.6.

3.1.2 ROC-test

The second type of tests by Meek and Hatfield [1994] are tests on the rate of change. In this case the difference between two consecutive elements of the vector is checked concerning limits. For a test with static limits, this can be phrased as:

$$f_i = ((x_i - x_{i+1}) > a_{max}) \vee ((x_i - x_{i+1}) < a_{min}). \tag{3.4}$$

Like for the LIM test, f_i is the resulting flag vector, x_i are the values of the dataset to be checked, and a_{max} and a_{min} are the maximum and minimum limits, respectively. In case of dynamic limits, two vectors have to be defined by the performing scientist. This can be phrased as:

$$f_i = ((x_i - x_{i+1}) > a_{max,i}) \vee ((x_i - x_{i+1}) < a_{min,i}). \tag{3.5}$$

The modification of the differences $x_i - x_{i+1}$ with a modifying function Λ as it was shown for the LIM tests is of course also possible for the ROC tests.

3.1.3 NOC-test

As a third type of test, Meek and Hatfield [1994] proposed a check on whether data does not change for more than a predefined number of values. This can be formulated as follows:

$$f_i = (x_i = x_{i-1} = ... = x_{i-n_{max}-1}). \tag{3.6}$$

It can be used to detect errors of the instrument, if it does not react to the environment anymore.
A useful enhancement for this test is a check on whether a certain number of consecutive error values are included in the dataset. These error values depend on the way the data is stored and read in by the software. Typical values are for example -9999 or not a number (NaN).[3]
The here presented tests will be used in an application in section 4.4.

3.2 Tests on statistical parameters

A usual approach to evaluate data is to make use of information on the statistical distribution of the data under consideration. Mainly statistical moments, like mean or variance (Tsay [1988]) or even higher ones (Vickers and Mahrt [1997]), are calculated and interpreted. This is also a usual way for change point models to analyse moments of data and in homogenisation as well. In this section several, different forms of quality checks, which base on information about the distribution of the data will be explained. Its focus is set to approaches, that use information on statistical parameters of a data vector. Since the procedures presented here will mainly be used for comparison to the method presented in section 3.3, examples are not given at this point. This will be done in an application in section 4.2, together with the tests.
This section begins with an overview on statistical parameters like moments and specific percentiles in section 3.2.1. Afterwards, four different tests based on this concept will be presented. These are the division of the dataset with a block window in section 3.2.2 and a sliding window in section 3.2.3. In the following, the method of trimmed moments in section 3.2.4 and in section 3.2.5 the bootstrapping of moments are shown.

3.2.1 Overview

In this section, information of the distributions from statistical parameters are used to perform quality checks. The parameters used for these tests can be divided into two main groups: the statistical moments and the parameters depending on quantiles.
The first group are the moments, both the standard and the centralised. If $f_X(x)$ is the probability density function (pdf), their definition is given by

$$\mu_k = \int\limits_{-\infty}^{\infty} x^k f_X(x)dx \tag{3.7}$$

for the standard kth moment and

$$\mu_{c,k} = \int\limits_{-\infty}^{\infty} \left(x^k - \mu\right) f_X(x)dx \tag{3.8}$$

for the centralised kth moment (Von Storch and Zwiers [1999], p. 32). With this it is possible to define for example the mean ($k = 1$), variance ($k = 2$), skewness ($k = 3$) and kurtosis ($k = 4$). For the mean, the standard and the centralised version is the same, for the higher moments exist two different versions.
The second group consists of the percentiles of the distribution. With the cumulative distribution function

[3] Of course other enhancements would not only concentrate on consecutive values, but check whether the number of any values exceeds a critical limit. This would be especially useful for the number of error values in a dataset.

(cdf) $F_X(x)$ the p-quantile x_p is defined by (Von Storch and Zwiers [1999], p.31)

$$F_X(x_p) = p. \tag{3.9}$$

In this context, the 0.5-quantile, also called median, indicates the value, where half of the values of the dataset are higher and the other half are lower. The 0.05- and 0.95-quantile deliver information about the values at the upper and lower tails of the distribution of the measurement vector. Besides this, other percentiles, like the 0.25- and 0.75-percentiles, might deliver useful information about the behaviour of the distribution of the vector.

For a proper analysis it is therefore necessary to take a look at several parameters, which lead to several plots and have to be evaluated. In section 4.2 an application will be shown, where this analysis can be simplified by the combination with other methods. The methods in the next sections specify the database for which these parameters are calculated. By comparing different databases for these calculations, it is possible to detect inhomogeneities within the dataset.

3.2.2 Block window

A first approach to calculate and visualise changes in the statistical parameters is the use of a block window. The dataset is divided into sections with a given length, specified by a parameter m_{block}. In each block the parameters are calculated separately. This delivers a result vector for every statistical parameter. Their length depend on the parameter m_{block} and on the handling of the potentially incomplete block at the end of the dataset. If the first block starts with the first element of the vector under investigation, the last block will only exist, if it has m_{block} elements. In other words it exists only, if the modulo of the vector and the parameter m_{block} is 0. Otherwise, the parameter of this last block can only be computed from a smaller number of elements. This might influence the results and therefore it is in some cases recommendable to leave the incomplete block out of the analysis in this type of test.

A usual application of a block-wise calculation of the statistical parameters in meteorology is the control of a time series, which includes diurnal or annual cycles. In these cases it is useful to choose sizes of a block, which include whole cycles and exclude incomplete blocks, to get a good representation of the dataset.

3.2.3 Sliding window

A similar approach to the block window is the use of a sliding window. Here, for the first section with length m_{slide}, the statistical parameters are calculated. Afterwards, the window is shifted forward iteratively, one element to another, and with each step the calculations are performed again. The result is a vector for every calculated statistical moment. They have a length of the original vector minus the parameter of the length of a section m_{slide} plus 1.

An advancement of this type of calculation window is the possibility to detect inhomogeneities with a precision of one element. In contrast to the block-wise method, the incomplete block at the end of the dataset is does not occur here. Nevertheless, the advice to cover included cycles within the dataset completely by a window stays in place. A consequence of applying this method is, that the number of calculations of the statistical parameters, which have to be performed, raise enormously. Furthermore, the results of the investigated sections are not independent, when the windows overlap.

3.2.4 Trimmed moments

Trimmed moments are only applicable to statistical parameters, which do not depend on percentiles. In a first step, the dataset is sorted by value and the parameters are calculated for the whole dataset. Afterwards, the dataset is divided into 100 blocks with equal length. In every step one block at the maximum and one at the minimum of the values is removed and the parameters are calculated again for the now trimmed vector.[4] This is done until the last two percents around the median of the vector are left. As a result, a vector with a length of 49 elements is generated for every calculated statistical parameter. The test is parameterless, since all facts necessary to perform it are known, if the measurement vector is given. By searching for breaks in the generated vectors, it is possible to get information on outliers within the datasets.

3.2.5 Bootstrapped moments

Bootstrapping, first introduced by Bradley Efron in 1979, is a method to estimate unknown distributions of data by resampling (Efron and Tibshirani [1993], p. 56). In order to use it for the estimation of the distribution of a statistical parameter, a vector with the length of the original vector is generated. It is a realisation of the same empirical distribution as the original dataset.[5] From the result, the statistical parameters are calculated and stored. This procedure is repeated with the number of repetitions, which is defined by a parameter r_{boot}. A possible way to use this procedure in a test, is to check the whole vector. If the results for a parameter have a large spread over the performed repetitions, it might be an indication of problems within the dataset. For example, if a large uncertainty is estimated for the quantiles at the tails, this can be an indication for outliers at the end of the distribution.

3.3 Histogram test

The histogram test is a new test to detect inhomogeneities in datasets and is described in Düsterhus and Hense [2012]. It does not only take the statistical parameters into account, but the whole distribution at once. This section will start with an explanation of the general methodology of the test in section 3.3.1. Therein, the necessity for distance measures of histograms will be emphasised. The five measures used in this thesis are presented and their calculation and characteristics are explained in section 3.3.2. To demonstrate the functionality, some sensitivity tests follow at the end in section 3.3.3. These tests show for example the performance of the recognition of shifts in the mean and variance within standardised vectors.

3.3.1 Methodology of the histogram test

The aim of the histogram test is to detect inconsistencies within a dataset. Therefore, the dataset will be divided into blocks with size s_b, which for one dimensional vectors is similar to the block window, described in section 3.2.5. In a next step every block is compared to every other block. This is done by comparing their normalised histograms, which are an estimation for the probability density function of the data within each block. The used number of bins of these histograms are defined as n_b. These bins are uniformly distributed between the maximum and minimum of both blocks, which are actually compared. The difference between two of these histograms is measured with distance measures, that will

[4] Especially for smaller datasets this might lead to errors, since the number of elements in each block varies. Reason for this is, that in most cases the length of the original vector cannot exactly be divided by 100. Therefore, it is possible that each package removed from the vector, has more or less elements than 1% of the original vector length.

[5] Technically this is done by a sampling with replacement from the original data vector.

be shown in the next section. The comparison of the histograms delivers one value for each comparison, that is afterwards stored in a result matrix. When the result matrix is filled, it is possible to detect inconsistencies within the dataset by looking for patterns in the matrix.

3.3.2 Distance measures for histograms

A usual field of application for the comparison of histograms is image retrieval (Rubner et al. [2001]). There, mostly multidimensional histograms are used and compared. In the application here, measures for one dimensional normalised histograms are required. Those are defined as $f \in \mathbb{R}^{n_b}$ and $g \in \mathbb{R}^{n_b}$. In the following, five difference measures are shown: Kullback-Leibler Divergence (KLD), Jenson-Shannon Divergence (JSD), Earth Mover's Distance (EMD), Root Mean Square (RMS) and Mean Square (MS).

3.3.2.1 Kullback-Leibler Divergence

The Kullback-Leibler Divergence was introduced by Solomon Kullback and Richard Leibler in their paper in 1951 (Kullback and Leibler [1951]). The definition used of the divergence in this thesis was given by Lin [1991]:

$$D_{KL}\left(f||g\right) = \sum_{i=1}^{n_b} f(x_i) \cdot \log_2 \frac{f(x_i)}{g(x_i)}. \tag{3.10}$$

The Kullback-Leibler Divergence D_{KL} uses a bin-wise comparison of the histogram and is no metric in the mathematical sense. Reason for this is the asymmetry of the divergence and that it does not obey the triangle inequality. The KLD is usually only defined, if both histograms have the same support, what means that both are positive definite for the same bins. This is especially required for the histogram g, because otherwise the denominator in the logarithm functions would become 0. To prevent this, prior information will fill all bins. The prior estimation, named a_p, is a uniform distribution on the whole domain of f and g. The value is added to every bin of both histograms before the comparison takes place. If h_i is a bin of the resulting histogram, the used equation is the following:

$$h_i = \frac{a_i + a_p}{s_b + n_b \cdot a_p}. \tag{3.11}$$

Here a_i are the number of observations in the bin i and s_b the total number of observations of the histogram. To define the prior a_p in more detail and make it scalable with the size of the blocks, the dependence is defined as follows:

$$a_p = \frac{1}{a_f \cdot s_b}. \tag{3.12}$$

The factor a_f has to be calibrated to the application. An example is given in section 3.3.3.2. Due to the introduction of this prior, the Kullback-Leibler Divergence is also defined, if f and g do not have the same support.

3.3.2.2 Jenson-Shannon Divergence

One disadvantage of the Kulback-Leibler Divergence is its asymmetry. This property was overcome by a symmetrisation of the KLD, which is known as the Jenson-Shannon Divergence (JSD). It is defined by Endres and Schindelin [2003] as:

$$D_{JS}(f||g) = \frac{1}{2} D_{KL}\left(f \left\| \frac{1}{2}\left(f+g\right)\right.\right) + \frac{1}{2} D_{KL}\left(g \left\| \frac{1}{2}\left(f+g\right)\right.\right). \tag{3.13}$$

Like the KLD, it is positive definite, but does not obey the triangle inequality. To become a metric, the root of D_{JS} has to be taken (Fuglede and Topsøe [2004]).

3.3.2.3 Earth Mover's Distance

The Earth Mover's distance (EMD) was developed by Rubner et al. [1998]. Unlike the other measures under consideration in this section, the EMD does not perform a bin-wise comparison of the two histograms. It is rather a solution to a transportation problem. The EMD is the minimised work required to transform one probability distribution to another (Levina and Bickel [2001]). The formulation for one-dimensional histograms can be given as (Rabin et al. [2008]):

$$D_{EM}\left(f||g\right) = \frac{1}{n_b}\sum_{i=1}^{n_b}\left|F_X(x_i) - G_X(x_i)\right|. \tag{3.14}$$

The measure compares the two cumulative distribution functions (cdf) of f and g, F and G, at every bin. It can be seen as the L1-Wasserstein metric (del Barrio et al. [1999], Levina and Bickel [2001]) and without the normalisation factor of n_b it complies with the match distance (Werman et al. [1985]). Unlike KLD and JSD, EMD is a metric in the mathematical sense (Rubner et al. [2000]).

3.3.2.4 Root Mean Square and Mean Square

The Mean Square (MS) and the Root Mean Square (RMS) measure is used as a reference in this thesis. The mean square is defined as:

$$D_{MS}\left(f||g\right) = \frac{1}{n_b}\sum_{i=1}^{n_b}\left(f(x_i) - g(x_i)\right)^2, \tag{3.15}$$

and the root mean square D_{RMS} is given by:

$$D_{RMS}\left(f||g\right) = \frac{1}{n_b}\left(\sum_{i=1}^{n_b}\left(f(x_i) - g(x_i)\right)^2\right)^{\frac{1}{2}}. \tag{3.16}$$

3.3.3 Sensitivity tests of the histogram test

In this section, some sensitivity studies demonstrate the functionality of the histogram test. It starts with the development of an evaluation method, that is used in the sensitivity tests explained in section 3.3.3.1. The first test is the determination of the prior information a_p and the value a_f, which defines the first in equation 3.12. This is necessary for the calculation of KLD and JSD. Afterwards, the detection quality of the different distance measures for level (section 3.3.3.3) and variance (section 3.3.3.4) shifts are evaluated and compared.

3.3.3.1 Evaluation method for the sensitivity tests

As a result, the method delivers a matrix, which includes the measured differences of the histograms. If the data is inconsistent and the distance measure is able to deliver indications for these inconsistencies, patterns will be detectable in the resulting matrix. A problem is to evaluate this detection in a way, that allows the performance of useful sensitivity studies.

The setup of all sensitivity studies for the histogram test is similar. It starts with a standard normal distributed vector of two thousand elements. This vector is divided into two parts. The first half is used

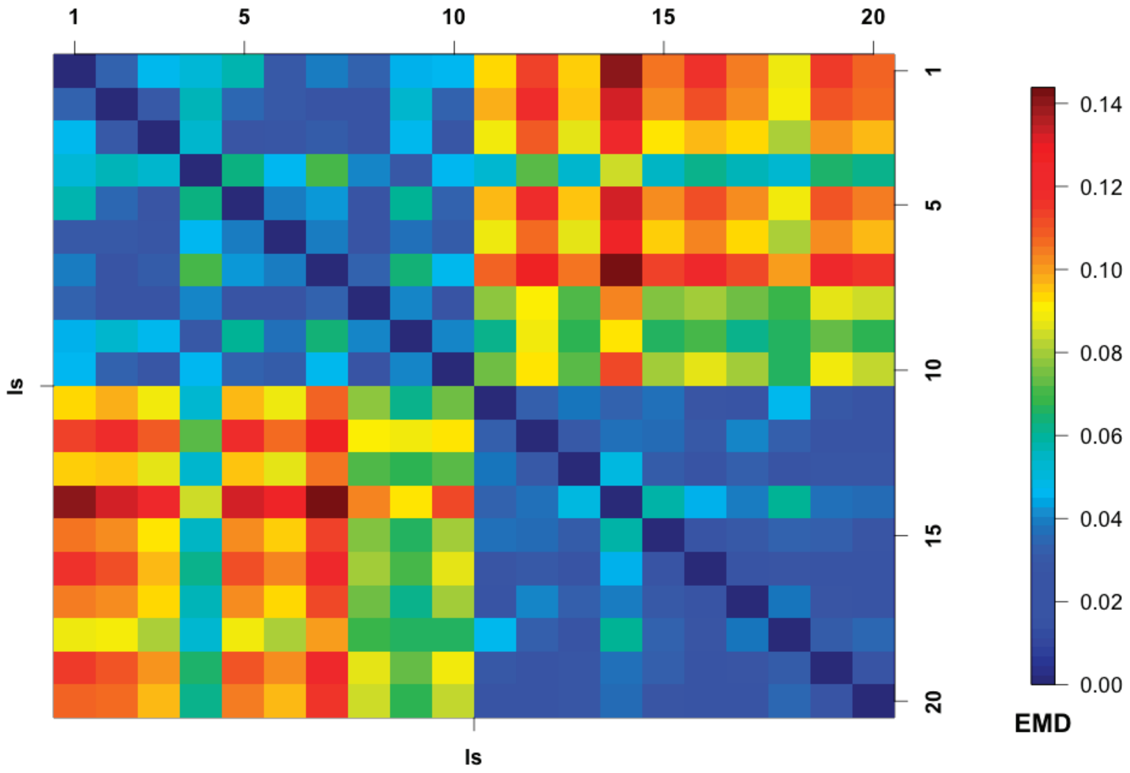


Figure 3.1: Example for the application of the histogram test with the distance measure EMD on a standard normal distributed vector with an included step at the middle of 0.4 standard deviations.

as it is, while the second gets some modifications. These modifications will consist of a rounding to the first digit in section 3.3.3.2, an added value in section 3.3.3.3 and a multiplied value in section 3.3.3.4. In a next step, the method is applied to this vector with the different distance measures. The result is a matrix for each used distance measure, that will be analysed afterwards.

The evaluation of this result matrices is based on the knowledge of the structure of these matrices. An example for such a result matrix is shown in figure 3.1. It shows the result of the histogram test with the EMD distance measure. The tested vector consists of normal distributed values, where in the second half a value of 0.4 standard deviations is added. The used parameters are a size of blocks of $s_b = 100$ and a number of bins of $n_b = 65$. The reason for the choice of the latter parameter will be explained in the next section. On the axes the number of blocks are shown, while each element in the matrix represents the measured distance between the histograms of the dedicated blocks. On the diagonal, each block is compared to itself. Therefore, both histograms, used in the comparison are the same, and as a consequence the distance between them is 0. Apart from this is a pattern recognisable, that divides the matrix obviously into four parts. On the upper left, the comparison of the first half with itself is shown. Since for this part the underlying distribution of the vector under analysis is the same, the measured distance is relatively low. The same holds for the lower right part, where the blocks of the second half of the vector are compared to themselves. The two other parts in the upper right and the lower left,

consist of higher values. Here, the first half is compared to the second half of the vector, where both of the underlying distributions have a different mean.
To recognise the pattern, it is useful to know the difference between the parts with the relatively high values and the parts with the relatively low values. Both types contain some variance, so it would not be helpful to just compare the mean of the values of each region. Therefore, additionally to the mean, the standard deviation of both sections is used for the evaluation. The mean of the section with the relatively low values, is named μ_{same}, as the same underlying distribution is used. The same reason holds for the mean of the region with the relatively high values, which is named μ_{diff}, because different underlying distributions are used here. The standard deviations are named likewise σ_{same} and σ_{diff}. Values in the matrix, which are equal to zero will not be included into the calculation of the means and standard deviations.
To calculate the difference between two sections, the value x_{sd} is computed, which fulfils the following equation:

$$\mu_{same} \pm x_{sd} \cdot \sigma_{same} = \mu_{diff} \pm x_{sd} \cdot \sigma_{diff}. \tag{3.17}$$

The consequence is, that x_{sd} measures the difference in standard deviations of both regions. The resulting equation for x_{sd} can be obtained with the use of some algebra:

$$x_{sd} = \left| \frac{\mu_{diff} - \mu_{same}}{\sigma_{diff} + \sigma_{same}} \right|. \tag{3.18}$$

The quantity x_{sd} is a measure to distinguish the two regions and is therefore appropriate to evaluate the detection quality of the methods.
A remark is necessary about which regions are used for the calculation of x_{sd}. Since the aim is to detect patterns, it is useful to compare the lower left to the lower right section. For symmetric distance measures this is equivalent the combination of the upper right to the lower right part. For the asymmetric measures this is not the case. The only asymmetric measure under consideration here is KLD. Therefore, when the KLD is compared to the other measures, the highest value of x_{sd} of the comparison of the lower left to the lower right and the upper right to the lower right is used. The method will be used in order to evaluate the modified vectors in the upcoming sections. The first modification is a rounding in the dataset.

3.3.3.2 Determination of the prior a_p

The aim of this section is to determine a useful value for a_f in the prior a_p in equation 3.12. Therefore, the calculation of the histogram test with the KLD and JSD distance measures are performed with a variation of the value a_f. The setup of this test uses eleven values for a_f, which are distributed on a logarithmic scale between 0.5 and 10000. Additionally, the number of bins n_b, used to perform these tests, is varied in the range of 2 to 201. The used test vectors are constructed the way it was described above (section 3.3.3.1), with the application of a rounding to the first digit as a modification of the second half of the test vectors. For each combination, the resulting matrices are evaluated by calculating x_{sd} of the lower left and lower right section. The procedure is repeated with one hundred different test vectors.
The mean of the 100 vectors is shown as a result in figure 3.2. The upper plot shows the results for the KLD, in the lower for the JSD. On the x-axis the number of bins n_b is shown, on the y-axis the factor a_f. The resulting x_{sd} are presented in steps of 0.5 for each combination in colours from blue for low values to red for high values. In both figures, the structure of the results is similar. The main variation is caused by the number of bins of the histograms n_b. The choice of a higher n_b leads to a higher value for x_{sd}. For the variation in a_f, it is possible to conclude, that for values higher than $a_f = 100$, no real changes are recognisable for the here chosen number of bins. An appropriate value of x_{sd} can be chosen with $x_{sd} = 1$

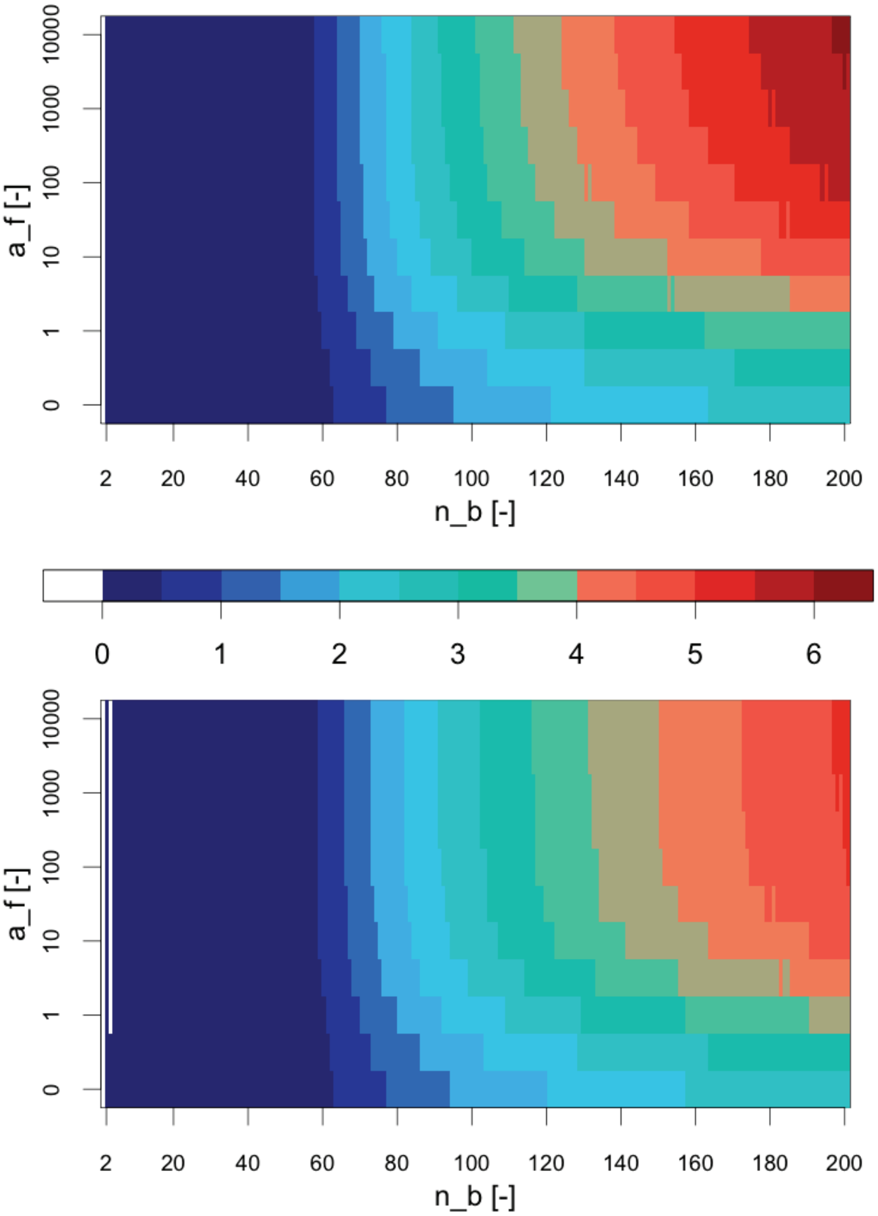


Figure 3.2: Sensitivity test of the histogram test with the distance measures KLD (top) and JSD (bottom) for rounding in data. On the x-axis the number of bins n_b is given, on the y-axis the prior a_f. Shown is the evaluation measure x_{sd}, averaged over 100 vectors.

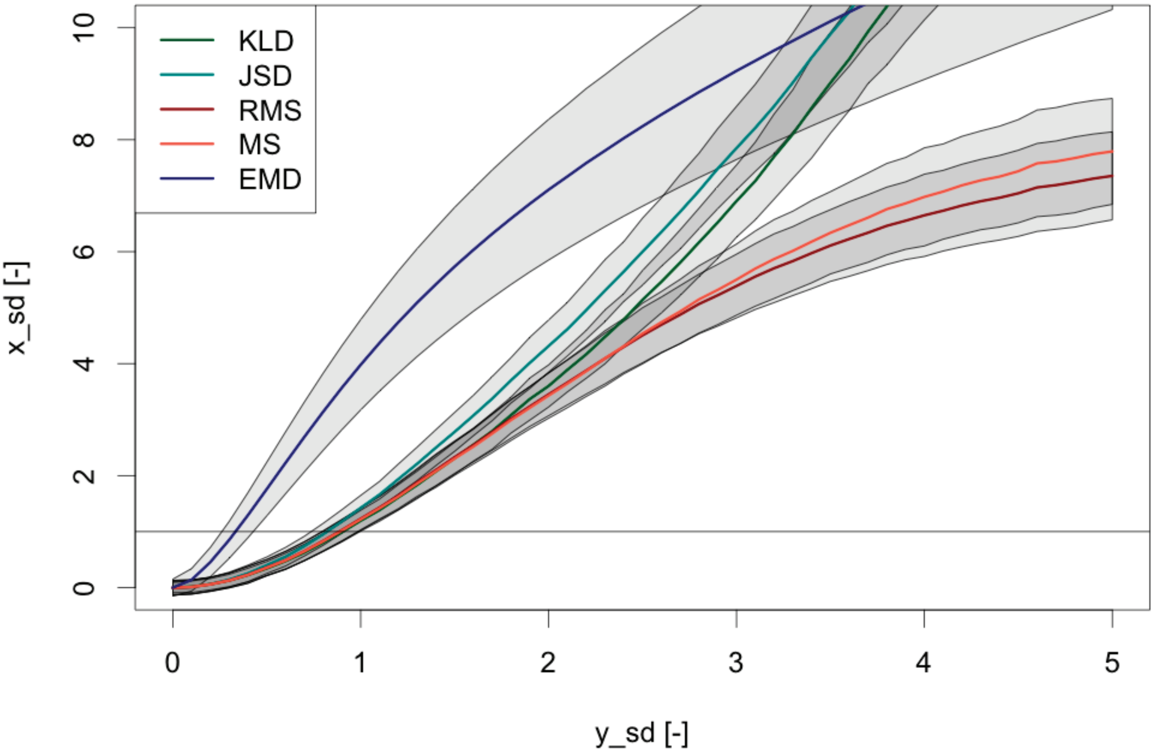


Figure 3.3: Sensitivity test of the histogram test for the five distance measures on detecting a shift in mean in standard normal distributed test vectors. On the x-axis the shift in the mean at the middle of the vector is measured in standard deviations y_{sd} and on the y-axis the evaluation of the result matrices is shown in x_{sd}. Results are averaged over 100 vectors, with the mean indicated by the horizontal line and the standard deviation indicated by the grey shadings behind the line.

for detecting a pattern. Therefore, an appropriate number of bins is around $n_b = 65$ in both cases. This value for n_b is used for all sensitivity tests and measures in the following. Further discussion and the interpretation can be found in section 5.1.1.1. In the next section, the modification of the underlying distribution of the vectors is a shift in the mean.

3.3.3.3 Shift in mean

The task of the next sensitivity test is to look for the distance measure that best detects level shifts within a dataset. Therefore, a value named y_{sd} is multiplied with the standard deviation of the original vector and added to the second half of the test vector. The quantity y_{sd} is varied in the range of 0 and 5. On each test vector and each added value the histogram test with all distance measures is performed. The results are shown in figure 3.3. On the x-axis the y_{sd} and on the y-axis the distinguishing measure x_{sd} is shown. The mean of one hundred test vectors for each distance measure is included as a line, the standard deviation as a grey shading behind.

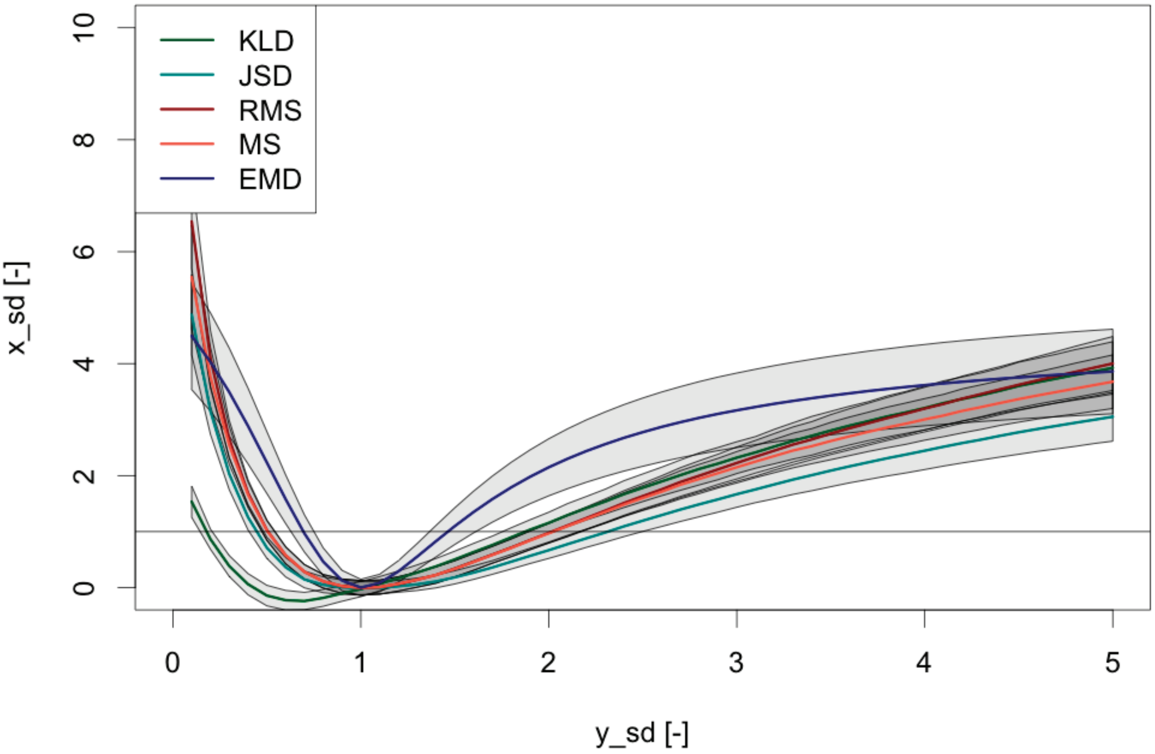


Figure 3.4: Sensitivity test of the histogram test for the five distance measures on detecting a shift in variance in standard normal distributed test vectors. On the x-axis the shift in the variance at the middle of the vector is measured in factors of standard deviations y_{sd} and on the y-axis the evaluation of the result matrices is shown in x_{sd}. Results are averaged over 100 vectors, with the mean indicated by the line and the standard deviation is indicated by the grey shadings behind the line.

For low y_{sd}, the results of the five methods split up into two groups. One of these groups only comprises the results of the EMD (blue) and increases much faster than the results for the other methods. The increase is nearly linear until 1.2 standard deviations. Afterwards, the slope decreases. The other methods, with a lower slope than the EMD, split up at a level shift of around $y_{sd} = 2.5$ standard deviations. The KLD (dark green) and JSD (cyan) increase further, while the MS (orange) and RMS (dark red) increase at a lower rate. For the detection of small steps in the test vectors, the EMD shows the best results. The vertical line, indicating the detection limit at $x_{sd} = 1$, is reached by the EMD at around $y_{sd} = 0.4$ standard deviations. The other methods reach this limit at about $y_{sd} = 0.9$. The difference between the two possibilities to evaluate KLD is only minor.

3.3.3.4 Shift in variance

As a last sensitivity test in this section, a shift in variance with a similar setup to the shift in mean is shown. The only change is applied to the term with the y_{sd}. This quantity is like before multiplied to the standard deviation, but then, instead of an addition, multiplied to the second half of the test vectors. The histogram test with the different measures is applied again to one hundred of such vectors, with the results presented in figure 3.4. Just like in figure 3.3, the y_{sd} can be found on the x-axis and the x_{sd} on the y-axis. The mean of the one hundred different test vectors for each distance measure is marked with a line. The colours are the same as those used in the last section. The standard deviation for the results of each method is shown again with the grey shading behind the line of the mean.

For a small increase in variance in the second half of the vector, the rise of the EMD delivers again the highest x_{sd} results. It reaches the detection limit of $x_{sd} = 1$ at around $y_{sd} = 1.5$. When the variance is decreased and therefore a y_{sd} lower than 1 is used, the EMD reaches again the detection limit first at around $y_{sd} = 0.7$. The other methods have, for an increasing variance factor y_{sd}, a nearly linear raise. Their slopes vary, but all methods reach the detection limit between $y_{sd} = 1.9$ and $y_{sd} = 2.4$. For decreasing the variance, JSD, MS and RMS behave similarly and reach the detection limit at around $y_{sd} = 0.5$. KLD reaches it at $y_{sd} = 0.2$. The KLD shown here is one of two options, that behave best for increasing the variance. The other option behaves better for a decreased variance, but much worse for an increased one. Consequences of the choice of the variance are discussed further in section 5.1.1.1.

3.4 change point test

In this section, a change point detection method will be introduced. The method is able to deliver probabilities on whether change points can be found in a time series. In section 3.4.1, it starts with an overview on change point detection methods. In the next section 3.4.2, the main change point detection method used here is explained. It was developed by Dose and Menzel [2004] and allows to regress different models to a dataset and to compare them by an estimation of their probability. An additional aim is to show different types of models, which can be regressed to datasets by the use of this method. In a next step, some sensitivity tests are performed with the change point detection method and its modifications in section 3.4.3. The results demonstrate the possible applications of the method and show which modifications are of practical use. Afterwards, the change point detection method by Dose & Menzel is compared to other methods in section 3.4.4. As a basis, the methods and tests, described by Ducré-Robitaille et al. [2003], are used. This section helps to classify the results delivered by the method of Dose & Menzel.

3.4.1 Overview

Change point tests and detection methods are common tools in statistical data analysis and are developed for many different purposes (see also for example Page [1955], Hinkley [1969], Hawkins [1977]). In the meteorological and climatological sciences they are mainly used in the field of homogenisation, where inconsistencies in time series are detected and corrected. For the detection exists an immense amount of different methods, of which some are presented in an inter-comparison study in section 3.4.4.1. Most of them regress the data and evaluate this regression in order to decide, whether a change point is found. If it is found, the method determines its location. In addition, also other methods, frequentist and bayesian, exist (Moreno et al. [2005]). Due to their large number they cannot be discussed here in detail. An overview of those used in homogenisation is given by the WMO in their "Guidelines on Climate Metadata and Homogenization" (Aguilar et al. [2003]). A prominent method, which is missing in the following inter-comparison study, is Caussinus and Mestre [2004]. They work with a penalised likelihood method and were not considered by Ducré-Robitaille et al. [2003]. Not discussed here are methods, that detect multiple change points, since most methods are developed to detect only one. Further explanations on this topic will be given in section 5.1.2.2.

In this thesis, the method developed by Dose & Menzel will be used and modified extensively, what will be described in the following section.

3.4.2 The method of Dose & Menzel

The change point detection method developed by Volker Dose and Annette Menzel was first presented in a publication in 2004. The statistical model was introduced and applied to find change points and trends in phenological data (Dose and Menzel [2004]). These studies, which used datasets up to one hundred data points, were extended to different phenological data and then used up to six hundred points in Menzel and Dose [2005]. In Dose and Menzel [2006] the method was also applied to seasonal averaged temperature data. Further studies in phenology, that base on this method were mainly compiled by Christoph Schleip, who performed intensive phenological studies for his PhD-Thesis under the supervision of Annette Menzel (Schleip [2009], Schleip et al. [2008], Schleip et al. [2009a], Schleip et al. [2009b]). Volker Dose used a similar method, based on the Poisson distribution, to fit change point models to hurricane data of the Carribean (Dose [2009]).
This section starts with a description of the model in section 3.4.2.1. Here, the equations and the basic framework for the method are given. The following section 3.4.2.2 explains the models, proposed by Dose and Menzel. In a last step, modifications of these models are presented and compared in section 3.4.2.3.

3.4.2.1 Theory of the method

The theory described in this section is based on the description in Dose and Menzel [2004]. In general, this method aims to fit a model described by the matrix $\mathbb{A}$ and functionals $\vec{f}$ to the data $\vec{d}$, which is available at timepoints $\vec{x}$. This can be written as

$$\vec{d} - \mathbb{A}\vec{f} = \vec{\epsilon}. \tag{3.19}$$

The error of the fit is described by $\vec{\epsilon}$, which is assumed to have a normal distribution with an expectation value of zero and a variance σ^2_{DM} ($\mathcal{N}(0, \sigma_{DM})$). To calculate the likelihood of the data under the condition of the described model with given $\mathbb{A}$ and $\vec{x}$, the following equation holds:

$$p(\vec{d}|\vec{x}, \mathbb{A}, I_B) = \int p(\vec{d}, \vec{f}, \sigma_{DM}|\vec{x}, \mathbb{A}, I_B) d\vec{f} d\sigma_{DM}. \tag{3.20}$$

Additionally to the background information, which will later be introduced in $\mathbb{A}$, some information, which is described by Dose and Menzel [2004] as "general conditional background" is introduced here as I_B. By using the product rule, this evolves to

$$p(\vec{d}|\vec{x}, \mathbb{A}, I_B) = \int p(\vec{d}|\vec{x}, \sigma_{DM}, \vec{f}, \mathbb{A}, I_B) p(\vec{f}, \sigma_{DM}|\vec{x}, \mathbb{A}, I_B) d\vec{f} d\sigma_{DM}. \tag{3.21}$$

The second probability under the integral, $p(\vec{f}, \sigma_{DM}|\vec{x}, \mathbb{A}, I_B)$, does not depend on $\vec{x}$ and $\mathbb{A}$. Therefore it is possible to split this term up to:

$$\begin{aligned} p(\vec{f}, \sigma_{DM}|\vec{x}, \mathbb{A}, I_B) &= p(\vec{f}, \sigma_{DM}|I_B) && (3.22) \\ &= p(\vec{f}|\sigma_{DM}, I_B) p(\sigma_{DM}|I_B). && (3.23) \end{aligned}$$

With the assumption, that $\vec{f}$ does not depend on the standard deviation σ_{DM}, there is now the necessity for two independent priors for both of them. For $\vec{f}$ a weakly informative prior is chosen with the introduction of the volume of a k-dimensional hypersphere V_s with a radius γ:

$$p(\vec{f}|\gamma, k, I_B) = \frac{\Gamma\left(\frac{k+2}{2}\right)}{\gamma^k \left(\sqrt{\pi}\right)^k} = \frac{1}{V_s(k,\gamma)}. \tag{3.24}$$

Both, γ and k, are taken from the general background I_B and Γ represents the gamma function. k is later chosen in dependence of the model, which is fitted to the data. An example, how k can be chosen, will be given in the next section 3.4.2.2. The influence of the parameter γ will be further analysed in section 3.4.3.

For the standard deviation σ_{DM}, an uninformative prior is chosen with a normalised form of the Jeffreys' prior (Berger [1985], p. 88):

$$p(\sigma_{DM}|\beta, I) = \frac{1}{2\ln\beta}\frac{1}{\sigma_{DM}} \tag{3.25}$$

Dose and Menzel [2004] chose the parameter β under the restriction $\frac{1}{\beta} < \sigma_{DM} < \beta$. To guarantee that condition, β is chosen in the methods presented here as follows:

$$\beta = \begin{cases} \sigma_{DM} + 1, & \text{if } \sigma_{DM} \geq 1 \\ 1 + \frac{1}{\sigma_{DM}}, & \text{if } \sigma_{DM} < 1 \end{cases}. \tag{3.26}$$

The second distribution in equation 3.21 is under the above mentioned assumptions for the error $\vec{\epsilon}$ given by:

$$p(\vec{d}|\vec{x}, \sigma_{DM}, \vec{f}, \mathbb{A}, I_B) = \left(\frac{1}{\sigma_{DM}\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma_{DM}^2}(\vec{d}-\mathbb{A}\vec{f})^T(\vec{d}-\mathbb{A}\vec{f})\right). \tag{3.27}$$

This equation can be transformed by introducing a matrix $\mathbb{Q}$ and the residual R to

$$p(\vec{d}|\vec{x}, \sigma_{DM}, \vec{f}, \mathbb{A}, I_B) = \left(\frac{1}{\sigma_{DM}\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2\sigma_{DM}^2}\left((\vec{f}-\vec{f}_0)^T\mathbb{Q}(\vec{f}-\vec{f}_0) + R\right)\right). \tag{3.28}$$

This transformation is later used to determine $\mathbb{Q}$ and R.

The aim is now to determine the probability $p(\vec{d}|\vec{x}, \mathbb{A}, I)$, which is given by equation 3.21. Therefore, Dose and Menzel [2004] performed the integration of equation 3.21 in two steps. The first is the integral over $\vec{f}$, which leads to:

$$p(\vec{d}|\vec{x}, \mathbb{A}, I) = \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \frac{1}{V_S(k,\gamma)}\frac{1}{2\ln\beta}\int_0^\infty d\sigma_{DM}\frac{1}{\sigma_{DM}}\frac{1}{\sigma_{DM}^2}\exp\left(-\frac{R}{2\sigma_{DM}^2}\right)\frac{2\pi\sigma_{DM}^2}{\sqrt{\det\mathbb{Q}}}. \tag{3.29}$$

Solving the remaining integral with respect to σ_{DM}, leads to the equation for the probability $p(\vec{d}|\vec{x}, \mathbb{A}, I)$ given by:

$$p(\vec{d}|\vec{x}, \mathbb{A}, I) = \frac{1}{2}\frac{1}{V_S(k,\gamma)}\frac{1}{2\ln\beta}\left(\frac{1}{\pi}\right)^{\frac{N-2}{2}}\frac{1}{\sqrt{\det\mathbb{Q}}}\frac{\Gamma(\frac{N-2}{2})}{R^{\frac{N-2}{2}}}. \tag{3.30}$$

The parameter N describes the number of elements of the time series $\vec{d}$.

To calculate this probability, it is necessary to specify the determinant of $\mathbb{Q}$ and the residual R. Dose and Menzel [2004] estimate both in dependence of $\mathbb{A}$ and $\vec{d}$. For further steps, the singular value decomposition

(Von Storch and Zwiers [1999], p. 415) of $\mathbb{A}$ is used:

$$\mathbb{A} = \sum_i \lambda_i \vec{U}_i \vec{V}_i^T. \tag{3.31}$$

The determination of $\mathbb{Q}$ and R is resulting from a comparison of coefficients of the transformation in equation 3.28. As a consequence, the following equation can be defined:

$$\mathbb{Q} = \mathbb{A}^T \mathbb{A}. \tag{3.32}$$

With the use of equation 3.31, it is possible to determine the determinant of $\mathbb{Q}$:

$$\det \mathbb{Q} = \Pi_k \lambda_k^2. \tag{3.33}$$

The residual R can be determined by

$$R = \vec{d}^T \left(\mathbb{I} \sum_k \lambda_k \vec{U}_k \vec{U}_k^T \right) \vec{d}. \tag{3.34}$$

Since both $\mathbb{Q}$ and R depend on $\mathbb{A}$, the latter has to be specified in the following. It will be shown, that with the variation of the composition of $\mathbb{A}$, it becomes possible to deliver an immense variety of different models, which can be evaluated by equation 3.30. Some examples are presented in the following, starting with those, that were used by Dose and Menzel [2004] themselves.

3.4.2.2 Different model types by Dose and Menzel

In Dose and Menzel [2004] the evaluation method to decide whether a time series is best fitted by a constant, linear or one change point model is developed. The constant model tries to minimise the error of the following model equation:

$$d_i - f = \epsilon_i. \tag{3.35}$$

To transfer this into the form given by equation 3.19, the matrix $\mathbb{A}$ simplifies to a vector of ones with the length of the data vector $\vec{d}$:

$$\mathbb{A}_C = (1)_i\,. \tag{3.36}$$

Another model usable is the linear model. Here, the model includes the opportunity of a trend. The basing model equation is given by

$$d_i - f_1 \frac{x_N - x_i}{x_N - x_1} - f_N \frac{x_i - x_1}{x_N - x_1} = \epsilon_i. \tag{3.37}$$

It uses the start and the end values (f_1 and f_N) of the functional, which ought to be regressed to the data. These values are weighted with the time elements x_i, that vary between the starting point x_1 and endpoint x_N.

To set up the matrix $\mathbb{A}$, the latter is divided into two parts, expressed by two columns:

$$\mathbb{A}_L = \left(\left(\frac{x_N - x_i}{x_N - x_1} \right)_i, \left(\frac{x_i - x_1}{x_N - x_1} \right)_i \right). \tag{3.38}$$

The last model, which is explained in full by Dose and Menzel [2004], is a generalisation of the linear

Table 3.1: Values for k for different models. n_{chp} describe the number of change points.

Model	k []
constant model	1
linear model	2
one change point model	3
n_{chp} change point model	$n_{chp} + 2$

model to a change point model. This model is defined by

$$d_i - f_k \frac{x_{k+1} - x_i}{x_{k+1} - x_k} - f_{k+1} \frac{x_i - x_k}{x_{k+1} - x_k} = \epsilon_i, \tag{3.39}$$

with $x_k \leq x_i \leq x_{k+1}$. Like before, this equation can be transformed to matrix notation, which leads to a $\mathbb{A}_{CHP}$. Since this procedure delivers large matrices, it will be shown graphically in comparison to other models in the next section. Additionally, the parameter k in equation 3.24 has to be set differently for all these models. Examples used by Dose and Menzel [2004] are given in table 3.1.

To evaluate, which model is the most probable, Dose and Menzel [2004] compared the different models by normalising the probabilities of each model. This can be calculated by the following equation:

$$p(\mathbb{A}_i|\vec{d}, \vec{x}, I) = \frac{p(\vec{d}|\vec{x}, \mathbb{A}_i, I)}{\sum\limits_{j \in \mathcal{J}} p(\vec{d}|\vec{x}, \mathbb{A}_j, I)}. \tag{3.40}$$

Here, $\mathbb{A}_i$ describes one model, for example the change point model $\mathbb{A}_{CHP}$. It is compared to other models, which are all included in the set $\mathcal{J}$. This might be set for example to $\mathcal{J} = \{\mathbb{A}_C, \mathbb{A}_L, \mathbb{A}_{CHP}\}$. Therefore, equation 3.40 calculates the relative probability of one model given a pre-defined set of models, including itself. This probability is used throughout the upcoming sections to compare different models. In the most cases, only the constant and change point model are compared. The only exception is the sensitivity test in section 3.4.3.1, where the linear model will be investigated as well.

A consequence of the change point model definition in equation 3.39 is, that two sections of the fitted model are always connected in one point. For some applications this might be useful, like Menzel and Dose [2005] and Dose and Menzel [2006] have shown. Nevertheless, for using this method in quality assurance applications, it is preferable to have a possibility to detect an instant step in the data. To achieve this, the change point model has to be modified, what is shown in the next section.

3.4.2.3 Modified model types

In the modification, a standard linear model equation is used, which leads to

$$d_i - a_k - b_k \frac{x_i - x_k}{x_{(k+1)} - x_k} = \epsilon_i, \tag{3.41}$$

with $x_k \leq x_i \leq x_{k+1}$. This model uses an intercept a_k and a slope b_k at every change point. The latter is multiplied with a quotient of time stamps, that start with zero at the beginning of the section. This implies, that only the intercept gives information on the value at this point. At the end of the section, this quotient becomes one, before the next section starts again by a point fully explained by the next intercept a_{k+1}. This delivers the asked possibility to have two independent sections in the model equation. The matrix $\mathbb{A}_{CHPm}$ for the change point model, based on the equation 3.41, will be shown graphically at the

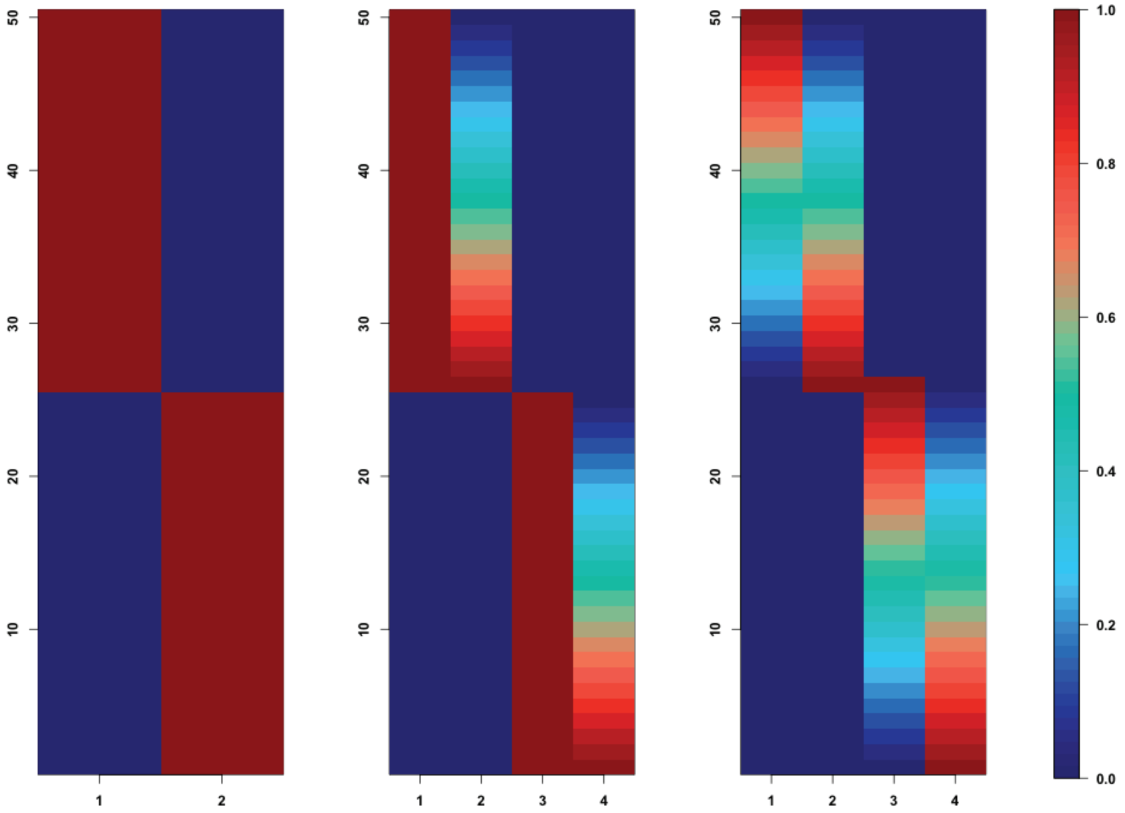


Figure 3.5: Matrices $\mathbb{A}$ for the different change point models of Dose-Menzel. On the left the matrix of the flat model, in the middle of the normal model and on the right of the matrix of the original model is shown. Colours range from zero (dark blue) to one (dark red).

end of this section. With this concept, it is also possible to modify the constant and the linear model. While the constant model is defined by

$$d_i - a = \epsilon_i, \tag{3.42}$$

and the matrix $\mathbb{A}_{Cm}$ is the same as $\mathbb{A}_C$ defined in equation 3.36, the linear model needs more modifications. Here, the model equation is given by

$$d_i - a - b\frac{x_i - x_1}{x_N - x_1} = \epsilon_i. \tag{3.43}$$

This leads to a matrix

$$\mathbb{A}_{Lm} = \left((1)_i \, , \left(\frac{x_i - x_1}{x_N - x_1} \right)_i \right). \tag{3.44}$$

Sensitivity tests and a comparison of both sets of models are shown in section 3.4.3.

As a third variant of the change point model, a version named 'flat model' will be used. It is similar to the modified change point model in equation 3.41, but sets the slope parameters b_k to zero. This implies the model equation

$$d_i - a_k = \epsilon_i, \tag{3.45}$$

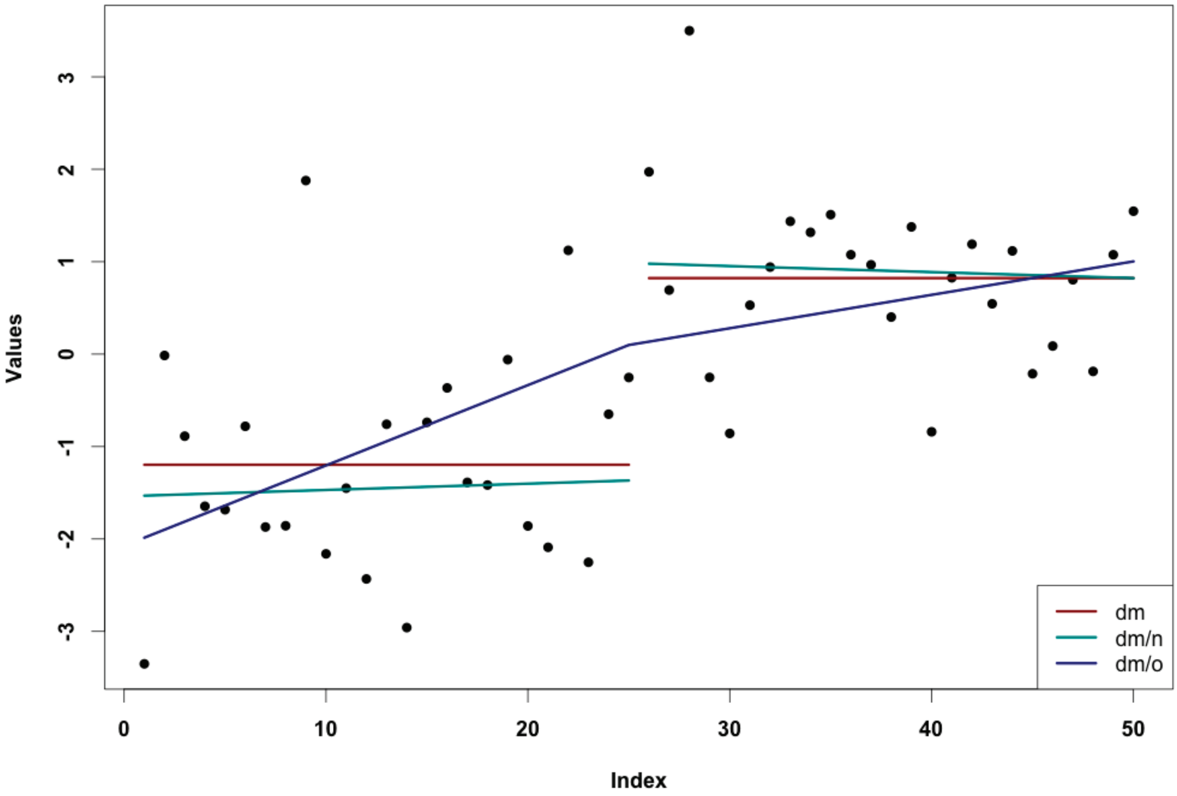


Figure 3.6: Application of the method of Dose-Menzel with the three different change point models to an artificial time series. The flat model is shown in red, the normal model in green and the original model in blue.

with $x_k \leq x_i \leq x_{k+1}$. This model uses two or more constant sections to fit to the data.

In figure 3.5 the three matrices ($\mathbb{A}_{CHPf}$ on the left, $\mathbb{A}_{CHPm}$ in the middle, $\mathbb{A}_{CHP}$ on the right) are compared. The values are taken from the intervall between zero in blue and one in dark red. They all show the situation for an estimated breakpoint at position 25 of a measurement vector with a size of 50. Therefore, they all have 50 rows. The flat model has only two columns for one breakpoint, the other two models four. The first column of $\mathbb{A}_{CHPf}$ shows entries with one up to the position of the estimated breakpoint, and zero afterwards. For the second columns it is the other way round. The matrix $\mathbb{A}_{CHPm}$ for the model given by the model equation 3.41 has four columns, which can be divided into two groups. Both groups consist of a constant value in the first part, which is filled with the value one up to the estimated breakpoint and zero afterwards. The second part is an increasing value from the beginning to the end of the section. The latter is marked by the boarder and the breakpoint. The last matrix, with the original model of Dose and Menzel consists of four columns that can be subdivided into two groups as well. The second part of each group is equal to its counterpart in the matrix $\mathbb{A}_{CHPm}$. The first part is a decreasing value from one to zero, whereas the second part increases.

To show the differences between the three model types, resulting from the matrices, an example is shown in figure 3.6. The three models are applied to an artificial time series with an included step. All three models detect the change point at the same position. The blue line represents the original model (dm/o) by Dose and Menzel. Here, the intercept and slope are regressed to the data within both sections, which are connected in one point. The same is regressed for the normal model (dm/n) with $\mathbb{A}_{CHPm}$, which is shown in green. The difference to the original model is, that the two sections are not necessarily connected at the change point. With the flat model (dm), shown in red, only the intercept is regressed and the two sections are not connected. After having presented the method and the different model, some sensitivity tests will show their abilities in the next section.

3.4.3 Sensitivity Tests and examples

In the following, some sensitivity test will demonstrate the advantages and disadvantages of the method by Dose and Menzel. Additionally, the influence of the different versions of the change point model, which were defined by the equations 3.39 for the original, 3.41 for the normal and 3.45 for the flat model, will be demonstrated.

This section consists of four sensitivity tests. The first in section 3.4.3.1 compares the constant, linear and one change point model for the three different formulations. The focus is set on the dependence on γ, which is the parameter of the prior in equation 3.24. In the following section 3.4.3.2, the dependence of the step detection is compared to the position and size of a step, which is included into a vector. Similar is the content of the next section 3.4.3.3. Here, the parameter γ and the size of the step are the varying parameters. In a last test in section 3.4.3.4, the parameter γ will be set for the following checks. The justification of this parameter depends on the probability that a step is detected in a homogeneous dataset.

The sensitivity tests are applied to artificial time series. They are chosen according to Ducré-Robitaille et al. [2003], what will be the basis for the inter-comparison of the Dose-Menzel method to other methods used in homogenisation. Their choice is an autoregressive process, which is from the statistical point of view (autoregression and variance) similar "...to those observed in the annual mean temperatures." The generating AR(1)-model is defined by:

$$X_i = 0.1X_{i-1} + \mathcal{N}(0,1). \tag{3.46}$$

In this case, X_i are the elements of a test vector and $\mathcal{N}$ is a normal distribution with the parameters mean and variance. The length of the selected vectors is 100 elements. It describes a normal distributed time series with an autocorrelation of 0.1. Additional tests with standard normal and gamma distributed time series will be shown in the inter-comparison tests in the sections 3.4.4 and 3.5.

3.4.3.1 Change point model comparison

The first sensitivity test investigates the behaviour of the three models for different values of γ, when they are applied to homogeneous data vectors. Homogeneous data means in this test, that the test is performed on time series generated by the process described in equation 3.46, without any artificially introduced steps. In the test 1000 vectors are used, of which each consisting of one hundred data points. The three models are applied to the vectors and the results, with the parameter γ varying between 1 and 40 are shown in figure 3.7.

Figure 3.7 consists of three identically structured plots. The upper one shows the results for the flat, the middle one for the normal and the lower one for the original model. On the x-axis the parameter γ is shown, on the y-axis the probability of each model. Each plot presents the averaged percentages of the change point model in green, the linear model in red and the constant model in blue. The standard deviation is marked as grey shadings behind the lines. Since only those three models are tested,

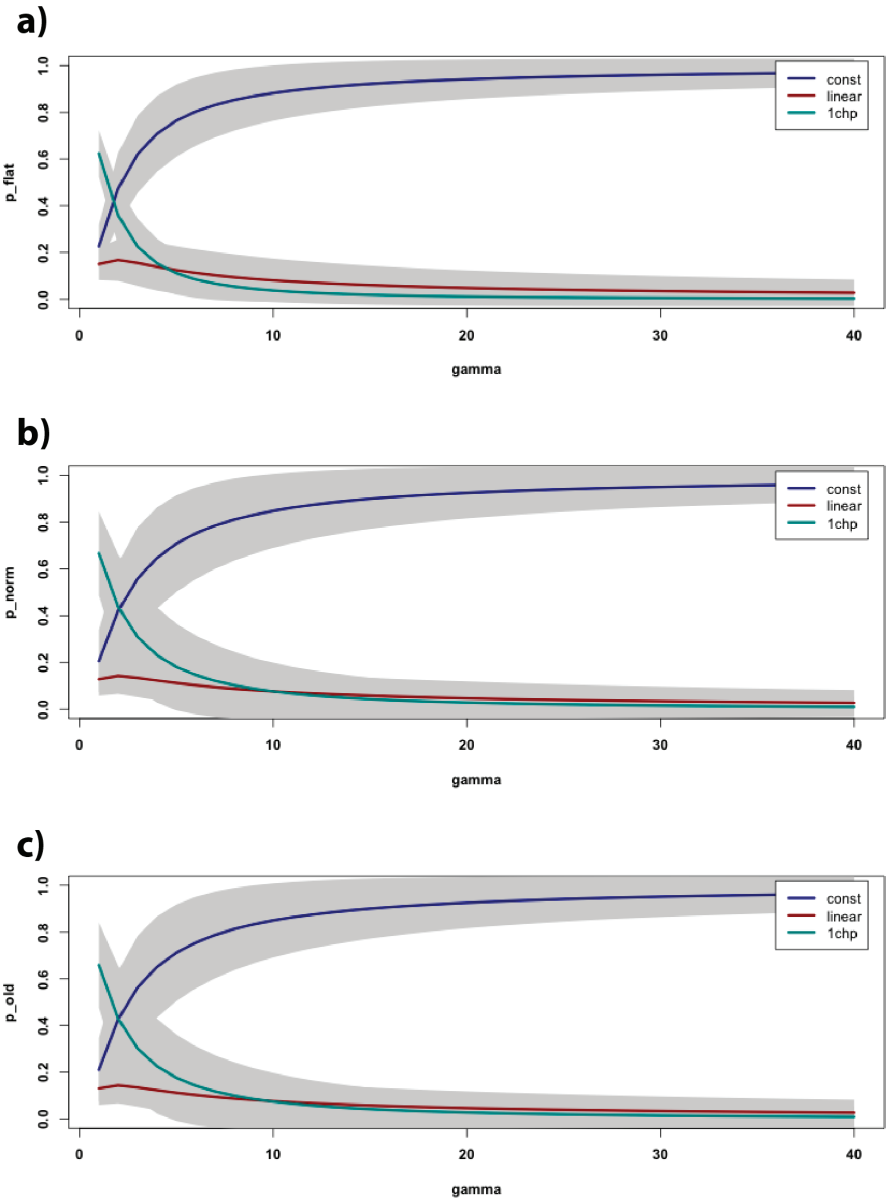


Figure 3.7: Application of the three types of models of the Dose-Menzel method to 1000 homogeneous vectors for varying γ. The x-axis illustrates the γ, the y-axis the probability of the methods. Relative probabilities for the flat model (top), normal model (middle) and original model (bottom) for the three used models constant (blue), linear (red) and change point (green) are averaged over the 1000 vectors. Grey shadings behind the lines indicate the standard deviations.

the probabilities sum up to one for every γ and formulation. The behaviour is similar for all three formulations: For low γ the change point model overtops the two others with a share of more than 60%. This changes, when the constant model gains more and more percentages, while the change point model declines. The linear model has only a low share for all γ, but is superior to the change point model for higher values of γ. The difference between the three formulations can be seen in the value of γ, where the turning points happen. While it is very similar to the normal and original model, the difference to the flat model is more visible.
The next tests use non-homogeneous test vectors with included artificial steps.

3.4.3.2 Position versus step

An important factor of a change point detection method is the dependence of the detection limit on the position of the step. The ideal case would be, that this limit is independent of the position of the step within the dataset. To demonstrate the performance of the three different methods in this sense, all were applied to one thousand different test vectors generated by equation 3.46 with a length of 100 data points. Each of these vectors is modified by the inclusion of artificial steps. These steps vary between 0 and 3 standard deviations and are included sequentially between all elements in the dataset.
The results of these tests are shown in figure 3.8. The upper left subfigure presents the result for the flat model (equation 3.45), the upper right for the modified model (equation 3.41) and the lower left for the original model by Dose and Menzel (equation 3.39). In each plot the percentage, where a breakpoint is detected with more than 95% probability, is plotted. These results are presented for the margin of the step and the first position, which is influenced by the step.
All three methods show, that their detection limit varies with the position of the step. For the modified and the flat model the behaviour does not depend on whether a step is included at the beginning or at the end of the dataset. The first 20 to 25 elements from the border of the dataset, the detection limit shows a strong decrease. Afterwards, the detection limit flattens for the inner steps. The original model by Dose and Menzel behaves similar. Nevertheless, with steps at the end of the vector the detection limit decreases much quicker than in the other methods. The difference in the results gets most obvious with a look at those positions and steps, where the methods detect inhomogeneities for at least 95% of the vectors under consideration. While the flat and the modified model are able to detect inhomogeneities up to the 96th position at a step of three standard deviations, it is much worse for the original method. The latter is only able to detect inhomogeneities for up to the 86th position. Since the minimum length of a linear section is chosen with three elements for all three methods, the modified and the flat models detect inhomogeneities for big steps at the borders of the datasets very well.
To investigate the optimal detection limit in the mid section of the vectors, the lower left plot in figure 3.8 compares the percentage of homogeneities for a fixed position and all used step sizes for all three methods. Therefore, the mean over the vectors is used, wherein the 50th element is the first data point, which is artificially modified with the step. The results with the flat model are shown in blue, with the modified model in green and with the original model in red.
All three lines rise steadily from smaller to bigger steps. They reach the five percent detection level of inhomogeneities at steps between 0.2 and 0.4 standard deviations and the 95-percent level between 1.2 and 1.4. For the flat model slightly better results than the other two approaches can be seen. The reason is that the detection limit is up to 0.1 standard deviations lower for the flat model. It is important here, to take into consideration, that the γ is 10 for all three models. As a consequence, it is impossible to say, whether the detection limit for the flat model is really better. In a next step, the behaviour of the different models, in respect to the variation of the parameter γ, applied to inhomogeneous datasets will be investigated.

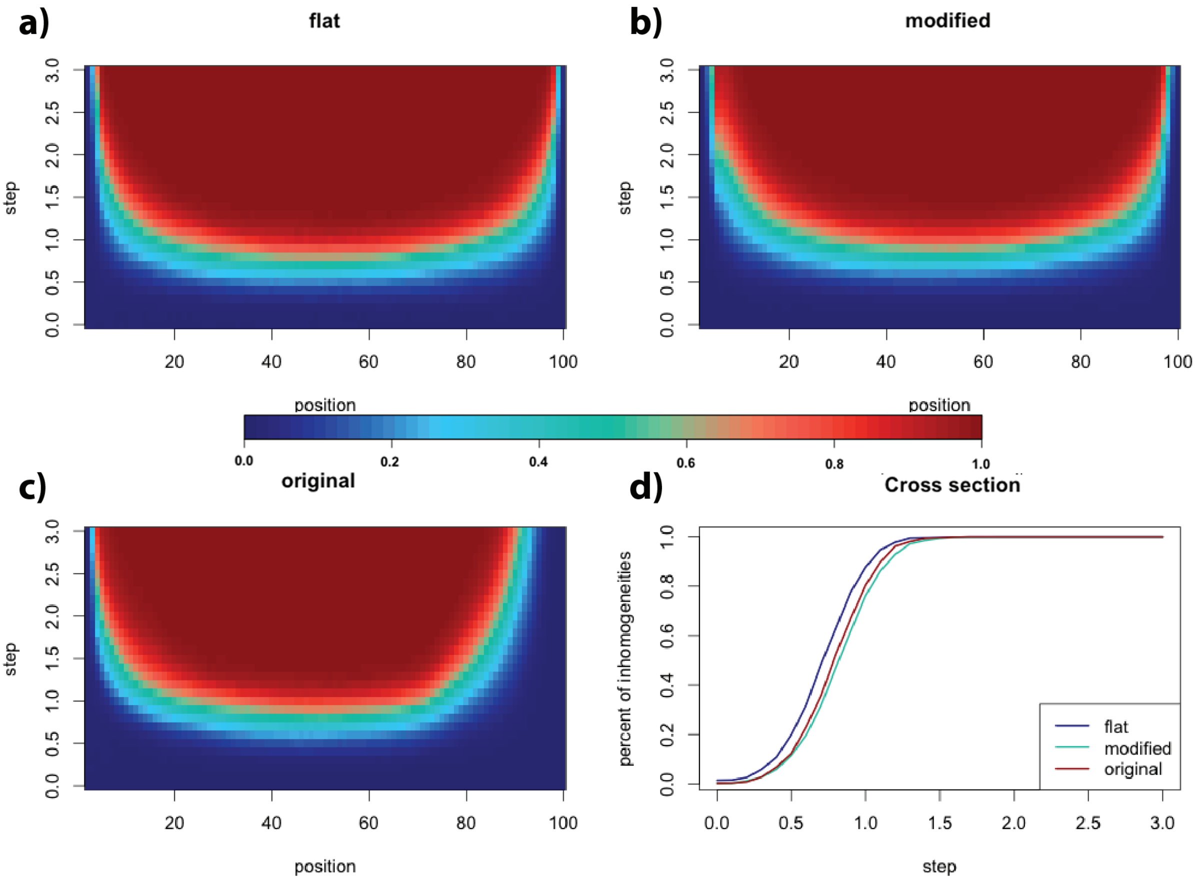


Figure 3.8: Sensitivity test on the dependence of the detection limit on the position and size of a step for the three modifications of Dose-Menzel. In the subfigures a-c, the x-axis indicates the position where the step is included and the y-axis illustrates the size of the step. Shown are the percentage of inhomogeneous vectors of 1000 vectors of the flat model (subfigure a), the normal model (subfigure b) and the original model (subfigure c). Subfigure d shows a cross section for the 50th position of the vectors. On the x-axis the step size is shown, on the y-axis the percentage of detected inhomogeneities.

3.4.3.3 γ versus step

The layout and the test settings of figure 3.9 are similar to the ones of figure 3.8. For this sensitivity test again one thousand different vectors, generated by the process described in equation 3.46, are investigated with the help of the three different models. In the upper left the results for the flat model can be found, in the upper right the ones for the modified model and in the lower right the ones for the original model by Dose and Menzel. The percentage of detected inhomogeneous datasets is shown on the x-axis for a variation of γ in the range of 1 and 40. On the y-axis the artificial steps at position 50, ranging between 0 and 3 standard deviations, are shown. Just like before inhomogeneity is assumed, when at any position the change point model has a 95%-probability level of superiority over the constant model.
All three show an increase of the detection limit with a rising γ. Additionally, all models show a similar rising of the slope, which is huger for smaller values of γ and more flatten for higher.
To further investigate the behaviour of the models the fourth plot in the lower left of figure 3.9 is used. It

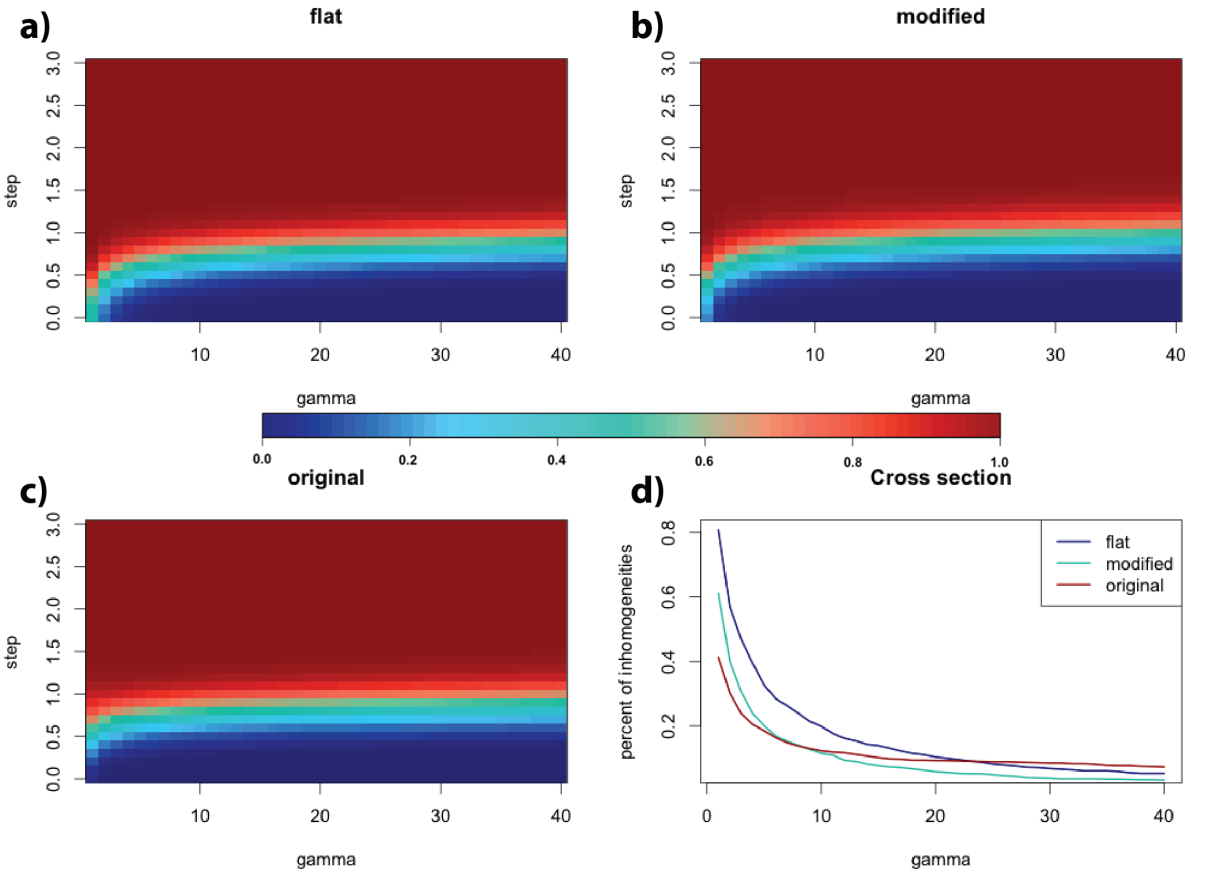


Figure 3.9: Sensitivity test on the dependence of the detection limit on the parameter γ and size of a step for the three modifications of Dose-Menzel. In the subfigures a-c, the x-axis indicate the γ and the y-axis the size of the step. Shown are the percentage of inhomogeneous vectors of 1000 vectors of the flat model (subfigure a), the normal model (subfigure b) and the original model (subfigure c). Subfigure d shows a cross section for a step of 0.5 standard deviations of the three former subfigures. On the x-axis, the step size is shown, on the y-axis the percentage of detected inhomogeneities.

contains a cross section, which shows the results for varying γ at a constant step with margin 0.5 standard deviations. This is obviously the transition zone of detection in the other three plots. The results from the flat model are shown in blue, with the modified model in green and with the original model in red. It is obvious, that the probability to detect an inhomogeneous dataset is declines for all three models with higher γ. For low γ up to about 20, the flat model has the highest probability of detecting a step, while the modified and the original model deliver similar results. For higher γ, the original model counts slightly the most detected inhomogeneous datasets.

The dependence on γ for the detection limit demonstrates, that this factor is a parameter, that can be used to justify the sensibility of the method. It also shows, that the detection of steps in the order of 1 to 1.3 standard deviations is possible. The flat version is the model with the lowest detection limits for most of the tested γ ranges. Especially the original method shows a flatter result, with higher detection limits for low γ and lower for higher γ.

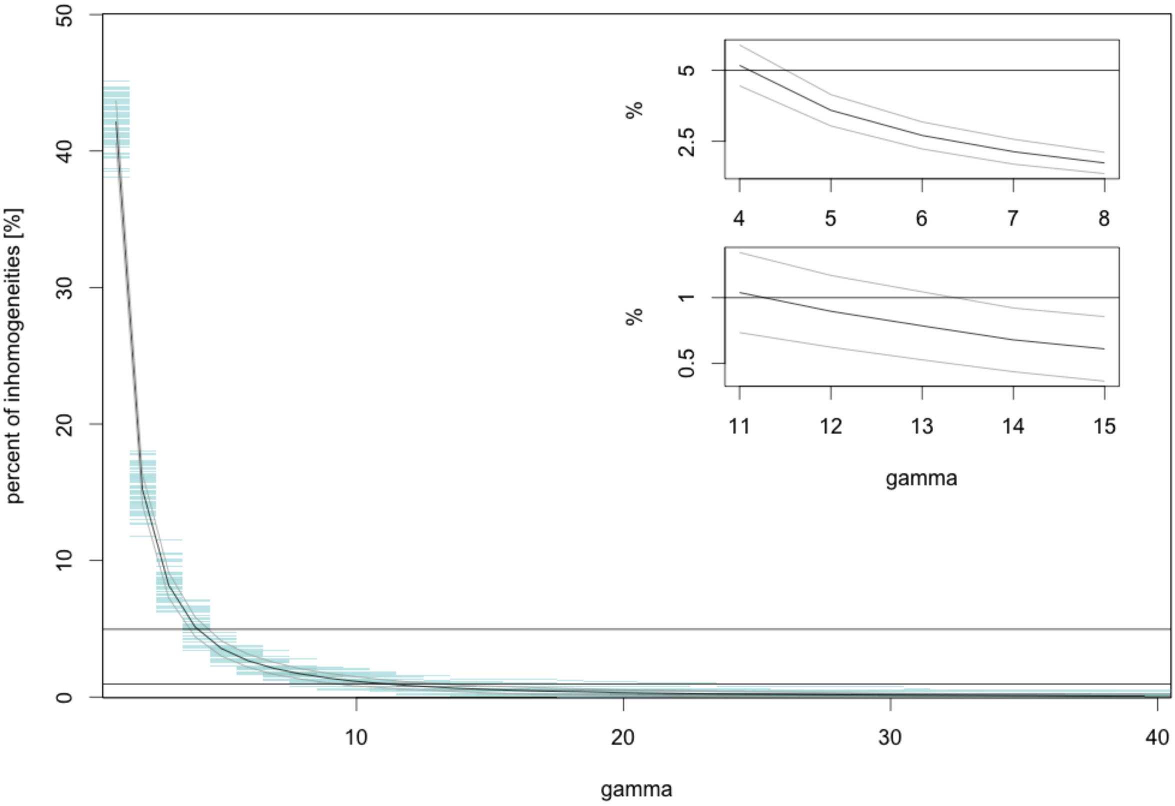


Figure 3.10: Justification of γ for the flat model with a 95% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.

3.4.3.4 Justification of γ

To justify the value of γ for the following comparison tests, a last sensitivity test is shown here. The three models test 100000 different homogeneous vectors generated by equation 3.46. The aim is to choose a γ, where the false positive rate is below a given threshold chosen here as 5%. The settings for the parameters are given like before: the smallest acceptable linear section is assumed by three elements ($m = 3$). A vector is assumed to be inhomogeneous, when the probability of a step at any position is more than 95%. The results of the 100000 vectors are split into one hundred packages of one thousand vectors each. The results of each package are averaged in order to demonstrate the uncertainty of the obtained significance levels.

In figures 3.10 to 3.12 the results for the three formulations are shown. Each figure shows the mean percentage of of all 100000 vectors as a black line. The grey lines indicate the standard deviation of the one hundred packages around the mean. The shaded lines in the background show each result of the packages separately. In the upper small figure on the right a focus is set on the five percent level of detected inhomogeneities, in the lower the focus is set on the one percent.

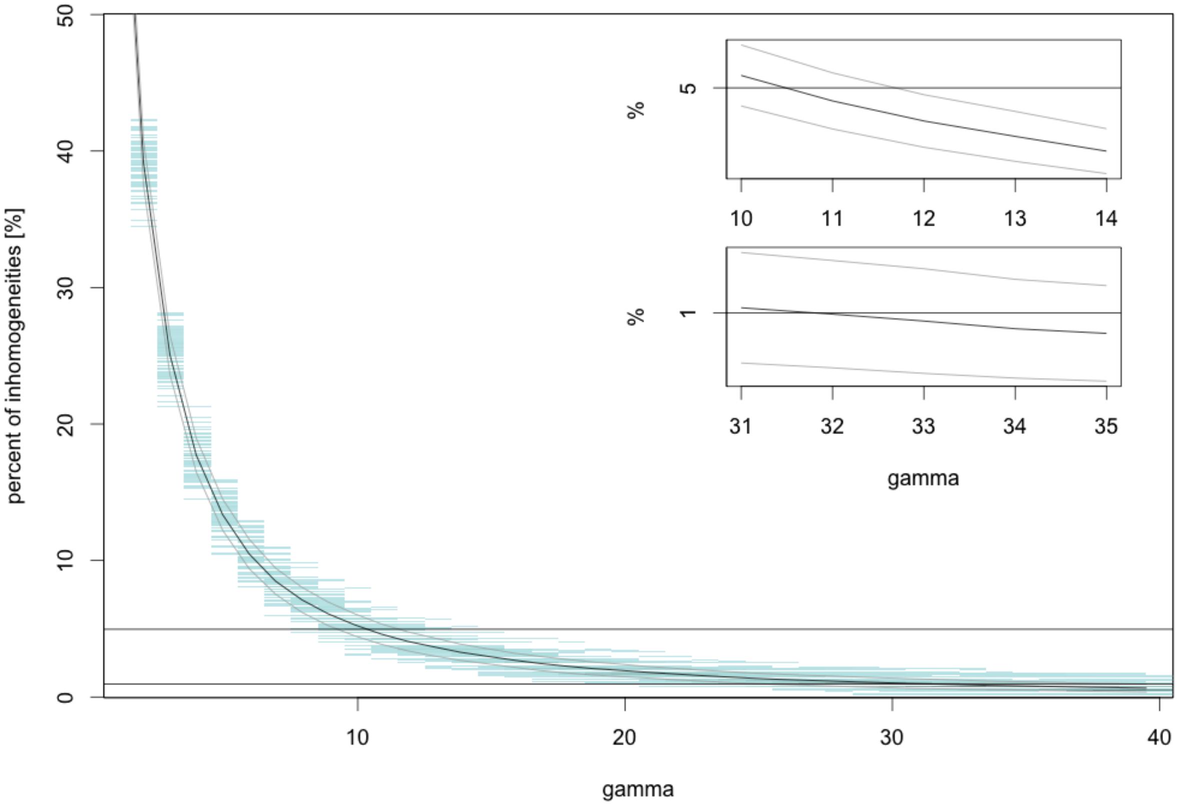


Figure 3.11: Justification of γ for the normal model with a 95% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.

In figure 3.10 the result of the flat model is shown. It shows a steady decline for increasing γ for the mean of the false positives. Also, the scattering of the results of the separate packages decreases. The threshold of 5% is reached by this formulation for $\gamma > 5$. For the lower threshold of 1% a $\gamma > 12$ is needed. The results for the normal method shown in figure 3.11 are structured similar, but reach the threshold for higher values of γ. The 5% threshold is reached for $\gamma > 11$ and the 1% threshold for $\gamma > 32$. Only slightly different to the normal method is the result for the original model, which is shown in figure 3.12. Here, the mean falls below the thresholds with $\gamma > 11$ for the 5% mark and with $\gamma > 21$ for the value of 1%.

For the comparison test in the following section 3.4.4 the highest values of γ are used, which fall short of the specified threshold. The same tests can be performed with different thresholds within the Dose & Menzel methods. Here, the results with a threshold of 95% are shown in the figures 3.10 to 3.12. In addition, the thresholds of 99% and 50% are used. For those, the same tests are performed and the γ is calibrated. Since the γ may exceed the maximal tested $\gamma = 40$, the maximal chosen value of γ is set to 40. For the threshold of 50% this might lead to a mis-calibration, but since the effect is minor it can be neglected. The results for the γ for all three threshold with a false alarm rate of 5% are shown in table

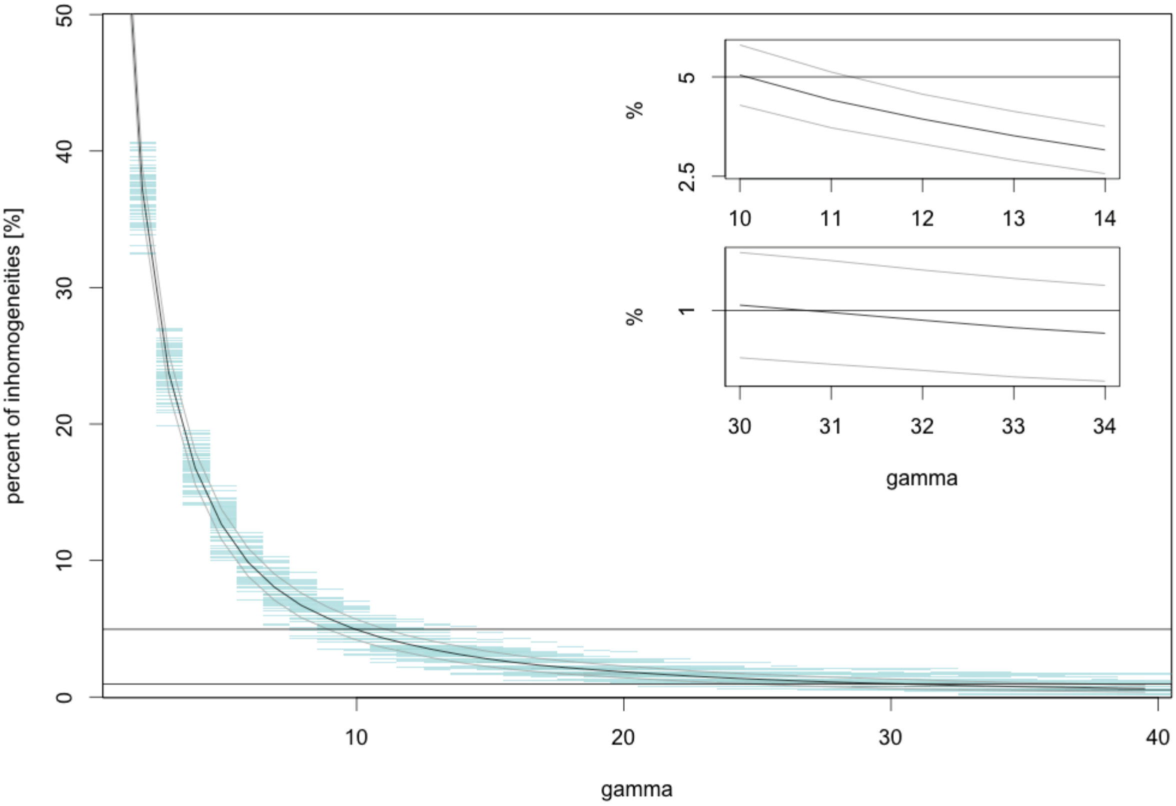


Figure 3.12: Justification of γ for the original model with a 95% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.

3.2.

3.4.4 Inter-comparison Tests

In section 3.4.2 the method by Dose and Menzel, its modifications and the setting of the parameters was described. They are now compared to other established change point detection methods in an inter-comparison study. For meteorological data, an immense amount of change point methods exists and some literature on the comparison of those methods has become available in recent years (e.g. Easterling and Peterson [1995], Peterson et al. [1998], Ducré-Robitaille et al. [2003], Rodionov [2004], DeGaetano [2006] and Reeves et al. [2007]). Applications of the change point detection methods can be found in homogenisation of temperature or precipitation time series, where the results of the detection methods are used to correct the data. For this field, Venema et al. [2012] delivers an inter-comparison study for monthly temperature series.

A problem for comparisons of change point methods is to find a common basis, because the detection methods are designed for different applications (Reeves et al. [2007]). To prevent problems with a different

Table 3.2: Values for γ for different models for a given threshold.

Model / Threshold	95%	99%	50%
flat model	5	2	18
normal model	11	5	40
original model	11	5	40

basis for the different tests, for the following comparison only methods from one study, Ducré-Robitaille et al. [2003], are used. They have chosen six different methods, wherein two methods are used in two different modifications. These eight procedures evaluate a time series, from now on called original series, with length N for a change point at position m_{chp}. Seven of them depend on a reference series. These reference series are common in the homogenisation environment, and are commonly defined as homogeneous time series, with the expectation of a high correlation to the original series (Peterson and Easterling [1994]).

In section 3.4.4.1, the methods, to which the Dose and Menzel procedures are compared, will be introduced. Afterwards, the test vectors as the basis of the here performed experiments, are described in section 3.4.4.2. With these vectors, two tests are performed in the following. A test with homogeneous datasets will be shown in section 3.4.4.3. A further test, that also comprises non-homogenous datasets follows in section 3.4.4.4.

3.4.4.1 Methods

The methods used in the inter-comparison tests were collected and described by Ducré-Robitaille et al. [2003]. They are listed together with their original source in table 3.3. In the following, these eight change point checks are described in detail.

a) TPR Two Phase Regression (TPR) was presented by Easterling and Peterson [1995]. The method uses a difference series, which is calculated from the difference between the original and a single reference series. In a first step, a linear regression model is fitted to this difference series. To estimate the goodness of fit the residual sum of squares between the difference series and the linear model ($RSS_{TPR,0}$) is calculated. A second step is used to take every datapoint of the difference series (m_{chp}) and to fit two regression models: One to the data before and the other to the data after the selected datapoint m_{chp}. For this, two phase regression model, the residual sum of square is calculated separately for both section and summed up ($RSS_{TPR,1}(m_{chp})$). To estimate the significance of a step at m_{chp}, the F-distributed test statistic by Solow [1987] is used:

$$U_{TPR,m_{chp}} = \frac{(RSS_{TPR,0} - RSS_{TPR,1}(m_{chp})) \cdot (N-4)}{3 \cdot RSS_{TPR,1}(m_{chp})}. \tag{3.47}$$

The parameters for the F-distribution, that are used to estimate the significance of an inhomogeneity at position m_{chp}, are the degrees of freedom given with 3 and $N-4$.

b) MLR Multilinear Regression (MLR) was developed by Vincent [1998]. It uses two regression models with autocorrelated errors e_i. The first is given by the following model equation, which comprises the original series x_{orig} and a reference series x_{ref}:

$$x_{orig,i} = a + cx_{ref,i} + e_i. \tag{3.48}$$

Table 3.3: Procedures used in comparison to the Dose-Menzel method.

Abbrevation	Method name	source
TPR	Two phase regression	Easterling and Peterson [1995]
MLR	Multi linear regression	Vincent [1998]
SNHTwoT	Standard normal homogeneity test without trend	Alexandersson [1986]
SNHTwT	Standard normal homogeneity test with trend	Alexandersson and Moberg [1997]
ST	Sequential testing for equality of means	Gullett et al. [1990]
WRS	Wilcox rank sum	Karl and Williams Jr. [1987]
BayeswoRef	Bayes method without reference	Ouarda et al. [1999], Perreault et al. [1999] and Perreault et al. [2000]
BayeswRef	Bayes method with reference	Ouarda et al. [1999], Perreault et al. [1999] and Perreault et al. [2000]

The autocorrelation in the error e_i is realised by:

$$e_i = \rho e_{i-1} + \mathcal{N}(0, \sigma^2). \tag{3.49}$$

Ducré-Robitaille et al. [2003] set ρ to 0.08. To estimate, if the original series is homogeneous, the Durbin-Watson test (Durbin and Watson [1950]) is used. It calculates the parameter D_{DW} with the equation:

$$D_{DW} = \frac{\sum_{i=2}^{N} (e_i - e_{i-1})^2}{\sum_{i=1}^{N} e_i^2}. \tag{3.50}$$

Limits for the significance of D_{DW} are given in Durbin and Watson [1951].
If the series is estimated as inhomogeneous, a second model will be fitted to the data for every possible change point m_{chp}. It uses the model equation

$$x_{orig,i} = a + b\mathbb{I}_{i \geq m_{chp}} + c x_{ref,i} + e_i, \tag{3.51}$$

with

$$\mathbb{I}_{i \geq m_{chp}} = \begin{cases} 1, & i \geq m_{chp} \\ 0, & i < m_{chp}. \end{cases} \tag{3.52}$$

For all fitted regressions, the residual sum of squares is calculated. The RSS of the second ($RSS_{MLR,1}(m_{chp})$) are compared with the RSS of the first model ($RSS_{MLR,0}$). The significance of the step is estimated by the following F-distributed test statistic:

$$U_{MLR,m_{chp}} = \frac{(RSS_{MLR,0} - RSS_{MLR,1}(m_{chp})) \cdot (N-3)}{(N-2) \cdot (N-3) \cdot RSS_{MLR,1}(m_{chp})}. \tag{3.53}$$

The parameters of the F-distribution are given by 1 and N-3.

c) SNHT The standard normal homogeneity test was developed by Hans Alexandersson. The original specification does not test for possible linear trends in the data (Alexandersson [1986]). It tests for every possible change point m_{chp}, if

$$U_{SwoT,m_{chp}} = m_{chp}\overline{z_{1:m_{chp}}}^2 + (N - m_{chp})\overline{z_{(m_{chp}+1):N}}^2 \tag{3.54}$$

exceeds a critical value. Therein, $\overline{z_{1:m_{chp}}}^2$ describes the mean of the difference series between the original and the reference series for the values before the possible change point. The same applies to $\overline{z_{m_{chp}:N}}$ for the values after m_{chp}.

A modified procedure was presented by Alexandersson and Moberg [1997] and includes a test for linear trends. Therefore, two positions are searched, that define the start (a_s) and the end point (a_e) of the section with a linear trend. The test value in this situation is defined by

$$\begin{aligned} U_{SwT,m_{chp}} =& - a_s\mu^2_{SwT,1} + 2a_s\mu_{SwT,1}\overline{z_{1:m_{chp}}} - \mu^2_{SwT,1}S_{SwT,e} - \mu^2_{SwT,2}S_{SwT,s} \\ &+ 2\mu_{SwT,1}S_{SwT,ze} + 2\mu_{SwT,2}S_{SwT,zs} - 2\mu_{SwT,1}\mu_{SwT,2}S_{SwT,se} \\ &- (N - a_e)\mu^2_{SwT,2} + 2(N - a_e)\mu_{SwT,2}\overline{z_{(m_{chp}+1):N}}. \end{aligned} \tag{3.55}$$

The expressions for $\mu_{SwT,1}$, $\mu_{SwT,2}$, $S_{SwT,s}$, $S_{SwT,e}$, $S_{SwT,se}$, $S_{SwT,zs}$, and $S_{SwT,ze}$ are given in the appendix in section A.2.1.

The critical values for $U_{SwoT,m_{chp}}$ and $U_{SwT,m_{chp}}$ are estimated in Alexandersson [1986] and Alexandersson and Moberg [1997]. They are "practically equal" (Alexandersson and Moberg [1997]). Therefore the values are taken from Alexandersson and Moberg [1997] and are interpolated to get the critical values for different N (see also the appendix in section A.2.2).

d) ST Sequential testing for equality of means by Gullett et al. [1990] uses the t-test

$$U_{ST,m_{chp}} = \frac{\overline{z_{(m_{chp}-N_2):m_{chp}}} - \overline{z_{(m_{chp}+1):(m_{chp}+N_2+1)}}}{\sqrt{\frac{\sigma^2_{(m_{chp}-N_2):m_{chp}}}{N_1} + \frac{\sigma^2_{(m_{chp}+1):(m_{chp}+N_2+1)}}{N_2}}}$$

for every potential m_{chp}. N_1 and N_2 are a defined number of points before and after m_{chp} and are set to 5. $\overline{z_{(m_{chp}-N_2):m_{chp}}}$ and $\overline{z_{(m_{chp}+1):(m_{chp}+N_2+1)}}$ are the mean on the difference series between the original and the reference series over the specified number of points before and after the potential step at m_{chp}. $\sigma_{(m_{chp}-N_2):m_{chp}}$ and $\sigma_{(m_{chp}+1):(m_{chp}+N_2+1)}$, are the standard deviations of the same sections.

Ducré-Robitaille et al. [2003] define with 5.8 a different threshold for the critical value of significance of an inhomogeneity than the normal t-statistics (2.3 for $\alpha = 0.05$). The values can be achieved by a change in the degrees of freedom of the student-t distribution from 3 to 1.1.

e) WRS Karl and Williams Jr. [1987] used a method to verify steps in time series, when metadata indicates them. It bases on the Wilcox Rank Sum (WRS) and with the modifications performed by Ducré-Robitaille et al. [2003], it is possible to use it in order to find the most probable step in a time series. First, the indices of the difference series between original and reference series are ranked by their values. The number of points before a possible step m_{chp} is called N_1 and N_2 afterwards. The ranks of the indices before and after the step are summed up separately ($S_{WRS,1}$ and $S_{WRS,2}$) and the lower of both is indicated by $S_{WRS,x}$. The corresponding length of the section is described by N_x. Now, the value

$$U_{WRS,m_{chp}} = \frac{S_{WRS,x} + 0.5 - \frac{N_x(N+1)}{2}}{\sqrt{\frac{N_1N_2(N+1)}{12}}} \tag{3.56}$$

is calculated for every change point m_{chp}. $U_{WRS,m_{chp}}$ can be approximated by a standard normal distribution (Ducré-Robitaille et al. [2003]), so the significance of a step can be estimated.

f) Bayes A Bayesian method is provided by Ouarda et al. [1999], Perreault et al. [1999] and Perreault et al. [2000]. Like Ducré-Robitaille et al. [2003], it is used on two different types of data: firstly on the original series itself and secondly on the difference series of the original and a reference series. The series under investigation, with length N, is divided into a section before a possible change point m_{chp} and after. In a first part, the probability of a position of a possible step m_{chp} is calculated by

$$p\left(m_{chp}|x_i\right) = \left(\frac{N}{m_{chp}(N - m_{chp})}\right) S_{Bay}(m_{chp})^{-\frac{N-2}{2}} \tag{3.57}$$

with

$$S_{Bay}\left(m_{chp}\right) = \left(\sum_{i=1}^{m_{chp}} \left(x_i - \overline{x_{1:m_{chp}}}\right)^2 + \sum_{i=m_{chp}+1}^{N} \left(x_i - \overline{x_{(m_{chp}+1):N}}\right)^2\right) \left(\sum_{i=1}^{N} \left(x_i - \overline{x_{1:N}}\right)^2\right)^{-1}. \tag{3.58}$$

$\overline{x_{1:m_{chp}}}$, $\overline{x_{(m_{chp}+1):N}}$ and $\overline{x_{1:N}}$ are the averages of the stated sections.
To decide, whether a series is homogeneous, it is necessary to calculate the probability of the size of a step (δ) at a given potential change point:

$$p\left(\delta|m_{chp}, x_i\right) = \frac{(N-2)^{-0.5}}{\sigma_{m_{chp}}(\delta) B\left(\frac{1}{2}, \frac{N-2}{2}\right)} \left(1 + \frac{\left|\delta - \left(\overline{x_{(m_{chp}+1):N}} - \overline{x_{1:m_{chp}}}\right)\right|}{(N-2)\sigma^2_{m_{chp}}(\delta)}\right)^{-\frac{N-1}{2}}. \tag{3.59}$$

B is the beta function and is given by

$$B\left(x, y\right) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \quad \text{with } \Gamma(x) = (x-1)!. \tag{3.60}$$

The variance is estimated by

$$\sigma_{m_{chp}}\left(\delta\right) = \frac{N S_{Bay}\left(m_{chp}\right)}{m_{chp}(N - m_{chp})(N-2)} \sum_{i=1}^{N} \left(x_i - \overline{x_{1:N}}\right)^2. \tag{3.61}$$

To decide, if a step is significantly different from 0, and as a consequence inhomogeneous, δ_l and δ_u are determined for a given significance level α by the following equation:

$$p\left(\delta_l \leq \delta \leq \delta_u\right) = \int_{\delta_l}^{\delta_u} p\left(\delta|m_{chp}, x_i\right) d\delta = 1 - \alpha. \tag{3.62}$$

The series is considered to be inhomogeneous if $0 \notin [\delta_l, \delta_u]$.

3.4.4.2 Basics of Inter-comparison experiments

In the following, some experiments will compare the methods of Dose & Menzel, its modifications and those described in Ducré-Robitaille et al. [2003]. The testing procedures used for this thesis also originate from Ducré-Robitaille et al. [2003]. In general, two tests are performed. The first check works on homogeneous time series. Here, the aim is to control the false alarm rate. In a second check, homogeneous and inhomogeneous datasets are controlled. In this case, the aim is to check the performance of every

Table 3.4: Properties of Dose and Menzel methods used in the sensitivity tests.

Abbreviation	Model	reference series	threshold
DM/ref	flat	yes	95%
DM99/r	flat	yes	99%
DM50/r	flat	yes	50%
DM/ref/n	normal	yes	95%
DM99/r/n	normal	yes	99%
DM50/r/n	normal	yes	50%
DM/ref/o	original	yes	95%
DM99/r/o	original	yes	99%
DM50/r/o	original	yes	50%
DM/noref	flat	no	95%
DM99/nr	flat	no	99%
DM50/nr	flat	no	50%
DM/nr/n	normal	no	95%
DM99/nr/n	normal	no	99%
DM50/nr/n	normal	no	50%
DM/nr/o	original	no	95%
DM99/nr/o	original	no	99%
DM50/nr/o	original	no	50%

method on detecting change points. At first, the test vectors used in both tests are described.
For the following tests, three different types of test vectors are used, each of which consists of 100 elements. The first was introduced by Ducré-Robitaille et al. [2003] and describes an autoregressive process, which was already defined in equation 3.46. Therefore, it is a standard normal distributed vector with an autocorrelation of 0.1. Ducré-Robitaille et al. [2003] also defined a dedicated reference series Y, based on the original vector:

$$Y_i = 1.5X_i + 0.1Y_{i-1} + \mathcal{N}(0,1). \tag{3.63}$$

The normal distribution $\mathcal{N}$ is again given as shown in equation 3.46, with the first parameter indicating the mean and the second the standard deviation.
The second test vector is standard normal distributed, without any autocorrelation. The original series is defined by:

$$X_i = \mathcal{N}(0,1). \tag{3.64}$$

The dedicated reference series is chosen in a way, that the difference to the original series is also a normal distributed vector, but has a much smaller standard deviation:

$$Y_i = X_i + \mathcal{N}(0,0.1). \tag{3.65}$$

As a third test vector, a gamma distributed time series is chosen. It is defined as:

$$X_i = \mathcal{G}(2,1). \tag{3.66}$$

The function $\mathcal{G}$ is the gamma distribution, wherein the first parameter defines the scale and the second the shape. Here, also the reference series is chosen in a way, that the difference series still have similar characteristics:

$$Y_i = X_i + \mathcal{G}(2,0.1). \tag{3.67}$$

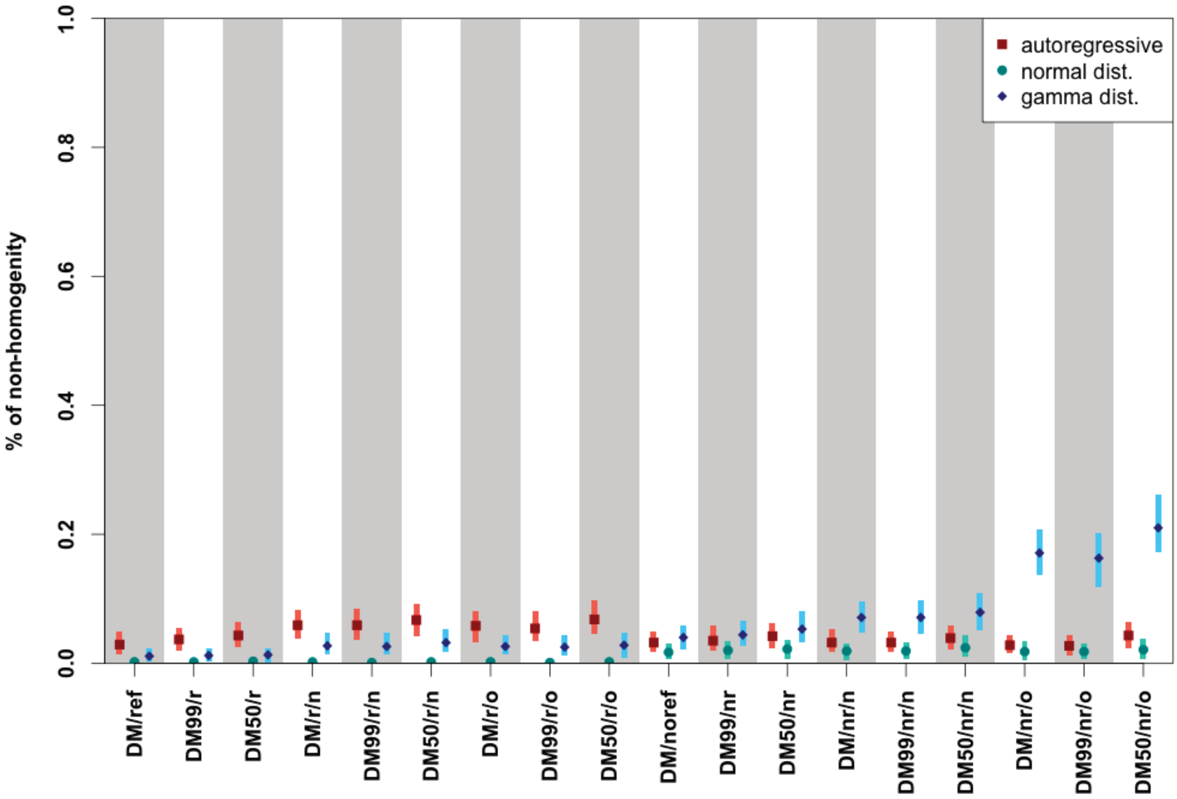


Figure 3.13: Application of the modifications of the Dose-Menzel method on homogeneous test vectors. On the x-axis the methods are shown, on the y-axis the percentage of detected inhomogeneities. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the value over 1000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

All three types of original and reference series are used in the following tests to demonstrate the influence of the basic distribution of the data on the outcome of the tests.

3.4.4.3 Test on homogeneous datasets

The first test is one on homogeneous datasets. The testing procedure used here was also applied by Ducré-Robitaille et al. [2003]. Unlike there, the focus here is not set on the height of the detected step, but on their existence. Therefore, 1000 vectors of each type are tested with each of the methods. Those are the eight methods, described by Ducré-Robitaille et al. [2003] and several combinations of the method of Dose & Menzel and its modifications. These combinations are collected together with their abbreviations in table 3.4. The abbreviations for the methods described in section 3.4.4.1 were already given in table 3.3.
The results of these tests are shown in the figures 3.13 and 3.14. The first, figure 3.13, will be used to describe the general structure of the plots, which will follow in a similar layout in this section and the next ones. The aim of the plot is to show the results of the different modifications of the Dose-Menzel

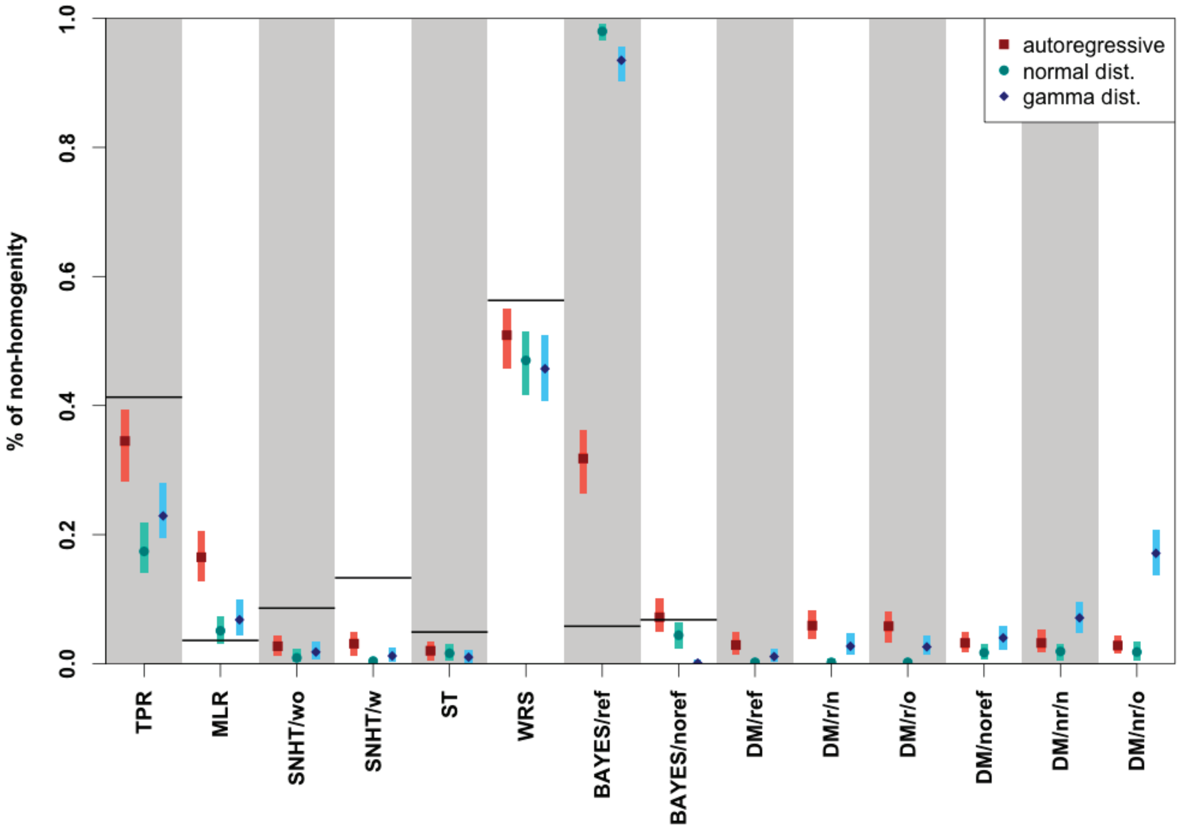


Figure 3.14: Inter-comparison test of several change point detection methods applied to homogeneous test vectors. On the x-axis the methods are shown, on the y-axis the percentage of detected inhomogeneities. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the value over 1000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples. Black lines indicate the results for the autoregressive test vectors calculated by Ducré-Robitaille et al. [2003].

methods. On the x-axis, the names of the methods are shown. The y-axis describes the number of detected inhomogeneous datasets in percent of the number of tested test vectors. For each method, the results are presented within a white or grey vertical stripe. In this, three markers with underlying bars are shown for each method. In red are the results of the autoregressive test vectors, in green the ones of the normal distributed test vectors and in blue are the results of the gamma distributed test vectors shown. The mark is the result for the 1000 vectors, while the bar shows the uncertainties as a bootstrapped estimation with 1000 samples (see also section 3.2.5).

A closer examination of the results of the different modifications of the change point detection procedure by Dose & Menzel shows, that their results are similar. The methods using a reference series can be found in the left half, without a reference series on the right. The three methods with different thresholds for a basic method are grouped together. For the autoregressive process in red all results are between 2.7 and 6.8%. The calibration of the methods has been performed on the autoregressive time series without using a reference series. Those results are in the range of 2.7 and 4.3%. This shows that the calibration has worked with an acceptable performance. For the results with the reference series, the inhomogeneities are a few percentages higher than without. The tests with the normal distributed test vectors deliver smaller percentages for all cases. While the methods without the use of a reference series detect some inhomogeneities, the others do not. A reason for this behaviour can be seen in the smaller variance of the time series under control for the methods with a reference series. The values for the methods with reference series for the normal distributed vectors range between 0.1 and 0.3% and without between 1.7 and 2.4%. The gamma distributed test vectors show a heterogeneous behaviour. For the methods with reference series the results are in the range of 1.1 and 3.2%. This means, that they are between the two other types of test vectors. Without the reference series, the gamma distributed vectors deliver the highest chance for the detection of inhomogeneities. Especially the original method by Dose & Menzel delivers high values up to 21%. Reasons can be found partly in the calibration procedure, which sets the maximal possible value γ to 40, even when higher values would be necessary. Nevertheless, only the method with the 50% threshold is influenced by this restriction. Still, the results of the original model with no reference series show, that they have to be taken with care in the following.

The second figure 3.14 has a similar structure. Here, the methods described by Ducré-Robitaille et al. [2003] are compared to the main methods from Dose & Menzel. The black lines for the first methods on the left indicate the results obtained by Ducré-Robitaille et al. [2003] in their analysis of the autoregressive process. Since the results of the methods are very heterogeneous, they are described separately. The first method on the left is the Two-Phase Regression (TPR) by Easterling and Peterson [1995]. It shows a high number of inconsistencies. The value for the autoregressive process determined here is at 34.5%. This result is in comparison to the value by Ducré-Robitaille et al. [2003] too low. Nevertheless, it is in the correct order, where the uncertainties of the bootstrap show, that it is only slightly out of range. The results for the normal and gamma distributed vectors are in the order of 20% (17.4% and 22.9%), what is lower than the autoregressive process, but still high. For the Multi Linear Regression (MLR), the results for the autoregressive process are by far too high (16.5 instead of 3.6%) in comparison to the results by Ducré-Robitaille et al. [2003]. The results for the normal and gamma distributed vectors are around 5% (5.1 and 6.8%).

The results for both of the SNHT methods are similar. Both deliver for the autoregressive process results at around 3% (2.7% and 3.1%). This is in the order of the 5%, what was the aim of the calibration by Ducré-Robitaille et al. [2003]. Still, they have much higher values (8.6% and 13.3%) as a result. For the normal and gamma distributed values, the results are very low in all cases. The ST method delivers for all three cases low values between 1.0 and 2.0 %. The original value by Ducré-Robitaille et al. [2003] is a little bit higher (4.9%). Also higher than the results in the experiments performed here are the results of the paper for the WRS method. In all cases, the results show a relatively high outcome. While Ducré-Robitaille et al. [2003] deliver a result of 56.3%, the autoregressive results are at 50.9% here. Normal and gamma distributed vectors are only slightly lower (47.0 and 45.7%). For the Bayesian methods, the results are very different. The result without a reference delivers a match between the results of Ducré-Robitaille et al. [2003] and the experiments performed here (7.2% to 6.8%). Contrastingly, the results with the use of the reference series are completely different. For the Bayes method with reference Ducré-Robitaille et al. [2003] determined the number of time series classified as inhomogeneous with 0.8%, while the autoregressive process delivers here 31.8% as a result. The results for the normal and gamma distributed vectors are even worse with 98.0% and 93.5%. For the Bayes method without a reference the result for the normal distributed vector is also in the range of acceptance with 4.4%. The gamma distributed vector leads to hardly any detected inhomogeneities (0.1%).

In an overview it can be said, that most of the reprogrammed methods deliver results, which are at least in the same order like Ducré-Robitaille et al. [2003]. Obvious exceptions are the MLR and the Bayes with

Table 3.5: Properties of Dose and Menzel methods used in the sensitivity tests.

		dataset	
		inhomogeneous	homogeneous
detection	inhomogeneous	$a = p_1$	$b = p_2$
	homogeneous	$c = p_3$	$d = p_4$

reference methods. Problems with those methods are briefly discussed in section 5.1.2.3. In comparison to the methods of Dose & Menzel, both SNHT methods, ST and Bayes without a reference series deliver similar results at around 5% for the autoregressive process.
In the next section, time series with included inconsistencies will be evaluated by all the above explained methods.

3.4.4.4 Detection of change points

The second test was also performed in a similar way by Ducré-Robitaille et al. [2003]. In this case, 25000 test vectors of each type of test vectors, described in section 3.4.4.2, are tested. The additional modification, which is applied to a time series, is the inclusion of steps. To include them, the explained algorithm is used in the following.
First, an element of the following exponential distribution $\mathcal{E}$ is sampled, where the parameter indicates the rate, which is set to 0.05:

$$\Delta t_{pos} = \mathcal{E}(0.05). \tag{3.68}$$

If the value for Δt_{pos} is greater or equal to 10, a step will be included at the Δt_{pos}-position after the last included step. If no step is included into the time series, it is the Δt_{pos}-position, where the step is included. If Δt_{pos} is lower than ten, a redraw takes place. Should a step be included at a position higher than 100, no additional step is added to the time series anymore. The size of the step is determined by the sampling of a standard normal distribution:

$$\Delta\delta_{step} = \mathcal{N}(0, 1). \tag{3.69}$$

If the absolute value of δ_{step} is in the range of 0.5 to 2, the step is included with this size at the given position. If not, the draw is repeated until a sufficient value is available. With these modifications, the test vectors are then checked by the methods. At this point, it is necessary to mention, that for the three types of vectors the same modifications take place.
The results of the methods, which are applied recursively to the datasets (see also section 5.1.2.2), deliver information about where the position of a change point is expected. In a first step of the analysis, the methods only use the information, if a step was detected or not by a method on a given test vector. With the knowledge about that and the information on the really included step modifications to the vector, it is possible to set up a contingency table. The one used is shown in table 3.5. The null hypothesis used, is that the tested vector is inhomogeneous.
To evaluate the contingency table, several scores are available in the literature. In the following, three different scores will be used. The first is the log odds ratio (Stephenson [2000]). It can be calculated by the following equation:

$$S_{lOR} = \ln\left(\frac{ad}{bc}\right). \tag{3.70}$$

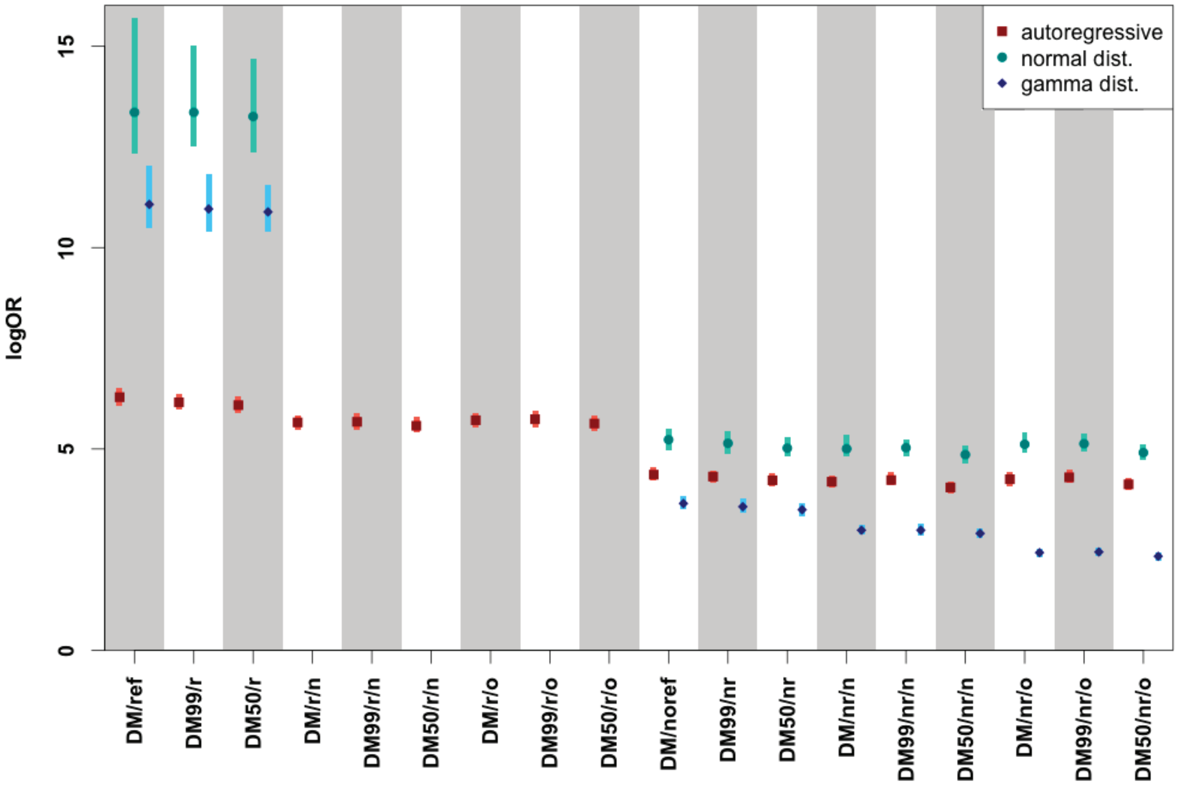


Figure 3.15: Log odds ratio of the application of the modifications of the Dose-Menzel method on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the log odds ratio. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

The odds ratio shows, how much better the model is in contrast to a random forecast. If both are independent, the odds ratio will become one and therefore the log odds ratio zero (Stephenson [2000]).
As a second score, the Heidke Skill Score (HSS) (Heidke [1926]) is used, which is defined by:

$$S_{HSS} = \frac{2(ad - bc)}{(a+c)(c+d)+(a+b)(b+d)}. \tag{3.71}$$

This score evaluates the number of hits (a) and correct rejections (d) and standardise them. The standardisation is constructed in a way, that a model which classifies the time series perfectly would lead to a S_{HSS} of 1. Generally, S_{HSS} can take the values between 1 and -1. It was also used by Menne and Williams Jr. [2005] to evaluate different change point detection methods.
The third method used is the calculation of the entropy. The equation can be defined by (Vigneron [2006]):

$$S_{Ent} = -\sum_i p_i \log_2 \left(\frac{p_i}{\sum_j p_j} \right). \tag{3.72}$$

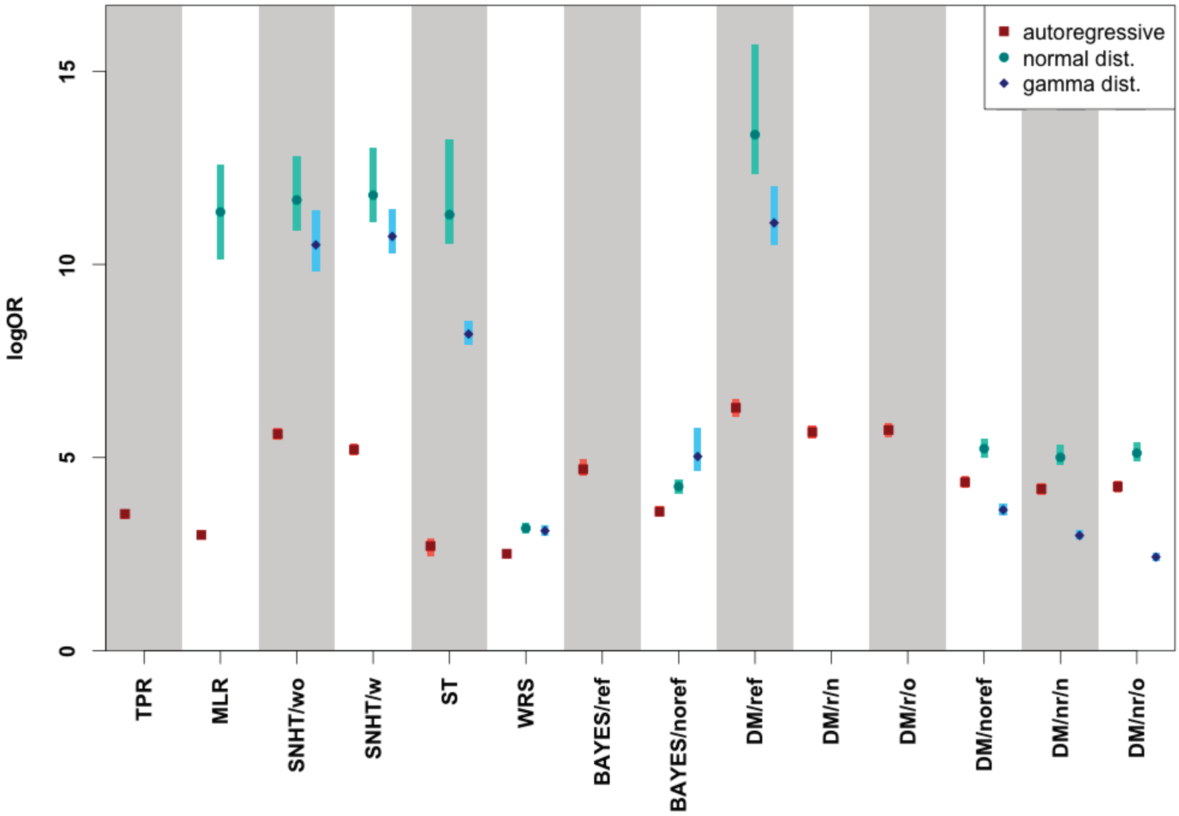


Figure 3.16: Log odds ratio of the inter-comparison of the modifications of the Dose-Menzel method and the methods described by Ducré-Robitaille et al. [2003] on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the log odds ratio. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

The theoretical limits for S_{Ent}, under the precondition that $0 \log_2(0) = 0$, are 0 and 2. The lower boundary indicates, that all weight is given to only one field of the contingency table. Reaching the upper boundary means, that all four entries are filled equally. In the following test, the lower limit is shifted upwards, due to the here given number of homogeneous (42,2%) and inhomogeneous (57,8%) test series. As a result, the lower theoretical limit is given by 0.983.

a) Log odds ratio At first, a look at the results of the log odds ratio will be taken. The analysis of this score starts with the different modifications of the method by Dose & Menzel, which are shown in figure 3.15. Like all the following figures in this section, it is similarly structured to figure 3.13, that was described in the last section 3.4.4.3. On the y-axis the log odds ratio is shown. Equally to the results for the tests on homogeneity, the main difference can be found in this plot for the methods, which use a reference series (right half) and those which do not (left half). For each type of test, the results are similar. The methods that use a reference series have a log odds ratio of around 6 for the autoregressive test vectors. For the normal and gamma distributed vectors ratio can only be calculated for the flat

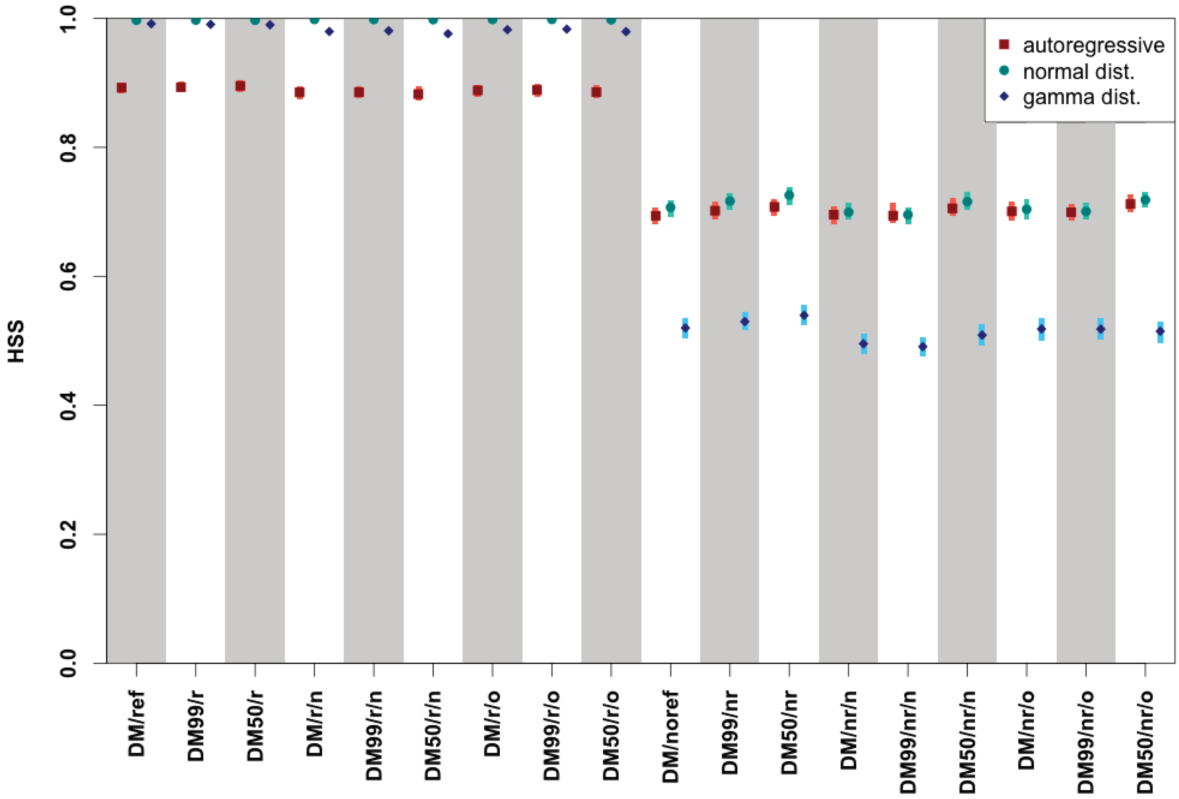


Figure 3.17: Heidke Skill Score of the application of the modifications of the Dose-Menzel method on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the Heidke Skill Score. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

model. Still, the uncertainties are also very high. This indicates, that the contingency table is not filled adequately for any of those applications. As a consequence, even the high values for the flat models cannot be used as a statement for a well performing model. For the methods without a reference series the results are lower, but the contingency table has enough entries in all fields. Here, the autoregressive test vectors deliver a log odds ratio of around 4, for the normal distributed ones of around 5 and for the gamma distributed vectors of around 3. This shows that the methods with reference, have results with higher values than those without a reference. Furthermore, these values are clearly higher than 0.

In a second step, the main modifications of Dose & Menzel are compared to the results described by Ducré-Robitaille et al. [2003]. On a first glance, several methods have the same problem with underdetermined contingency tables for the normal and gamma distributed test vectors, like the Dose-Menzel methods with reference. MLR, both SNHT modifications and ST show a high uncertainty for these test vectors as well. Therefore, the focus will be set on the time series, that are produced by the autoregressive process. Here, the highest values from the methods by Ducré-Robitaille et al. [2003] have comparable results to the modifications with a reference series from Dose & Menzel. From the first, both SNHT methods have

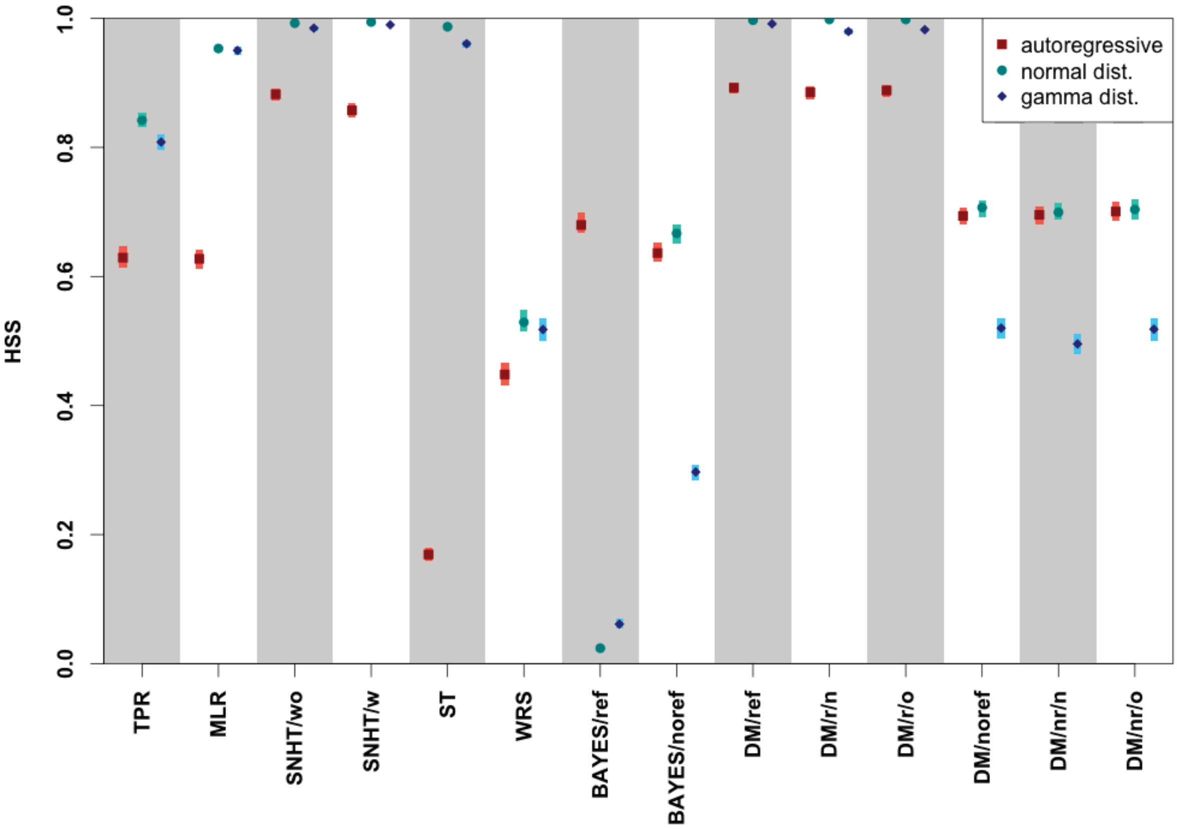


Figure 3.18: Heidke Skill Score of the inter-comparison of the modifications of the Dose-Menzel method and the methods described by Ducré-Robitaille et al. [2003] on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the Heidke Skill Score. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

results of 5.6 and 5.2, respectively. Like mentioned before, the latter reaches values between 6.1 and 6.3. From the other methods, only the Bayes methods with and without a reference series and TPR, with a log odds ratio of 4.7, 3.6 and 3.5, are in the range of the Dose-Menzel methods without a reference. The others show lower scores.

b) Heidke Skill Score In a next step, similar plots like for the other scores are shown for the Heidke Skill Score (HSS). At first, the different modifications of Dose-Menzel are investigated, what is shown in figure 3.17. Here, the HSS is shown on the y-axis, while the rest of the plot is structured similarly to figure 3.15. The plots show similar results as well. The main difference can be found between the methods, which use a reference series and those that do not. For the methods with a reference series the results of normal and gamma distributed test vectors are above 0.97, while they are at around 0.89 for the autoregressive test vectors. The methods that do not use a reference series have results of the autoregressive and normal types of test vectors grouped together at around 0.7. The gamma distributed vectors deliver lower results at around 0.5. The uncertainties are low for all cases, but they are higher

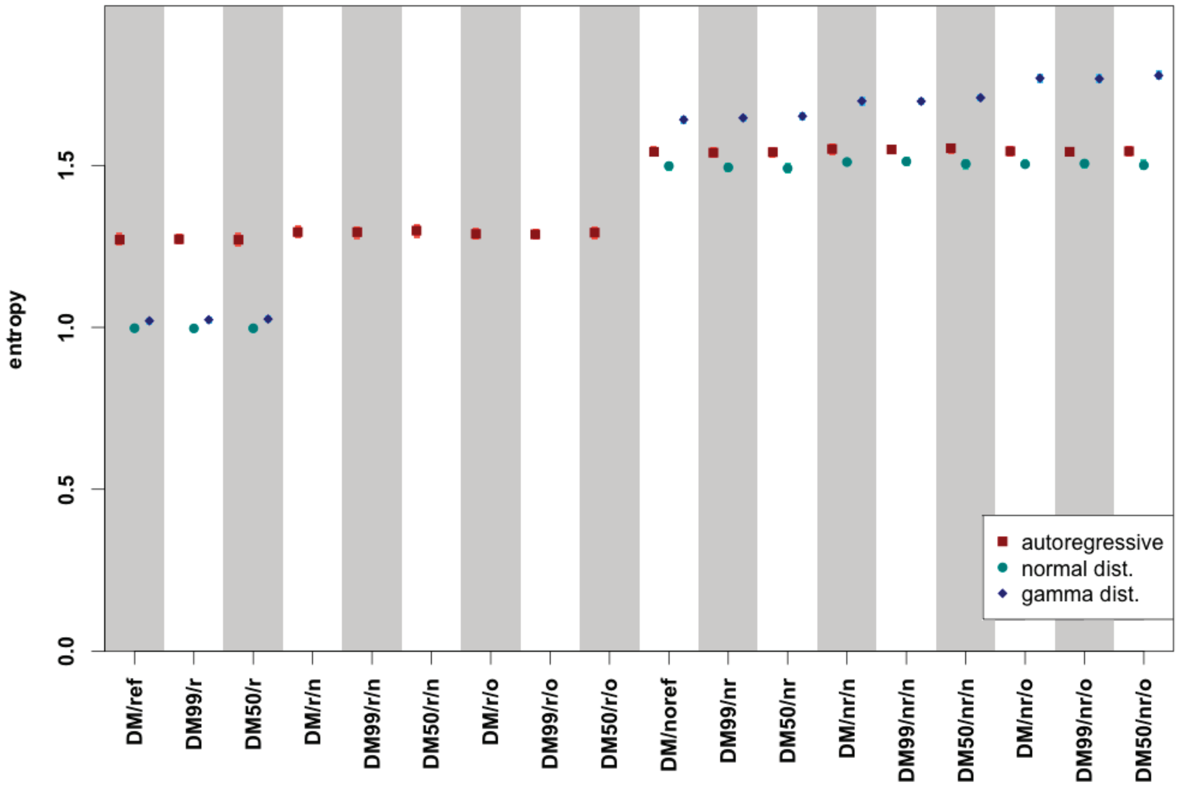


Figure 3.19: Entropy of the application of the modifications of the Dose-Menzel method on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the entropy. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

for those, which do not use a reference series than for those, which do.
In the case of the comparison of the methods described by Ducré-Robitaille et al. [2003] and the main modifications of Dose and Menzel the results, shown in figure 3.18, are more heterogeneous. Comparable results to the reference series using methods of Dose and Menzel are only reached by the two modifications of SNHT. ST has results on the same level as the SNHT methods for normal and gamma distributed vectors, but the autoregressive test vectors deliver results of only 0.17. The Bayes method without reference delivers similar results like the comparable methods of Dose and Menzel. TPR and MLR have high results for the normal and gamma distributed test vectors, while the autoregressive process shows results on the same level as the Dose-Menzel methods without reference series. WRS and Bayes with reference are hardly getting high scores at all. Only the latter shows values for the autoregressive process, which are compatible with the Dose-Menzel methods without reference series.

c) Entropy Finally, a look will be taken at the results of the entropy measure, which is given by equation 3.72. The comparison of the different modifications of the Dose-Menzel method is shown in figure 3.19. The results are again split between modifications using a reference series and the ones without. The

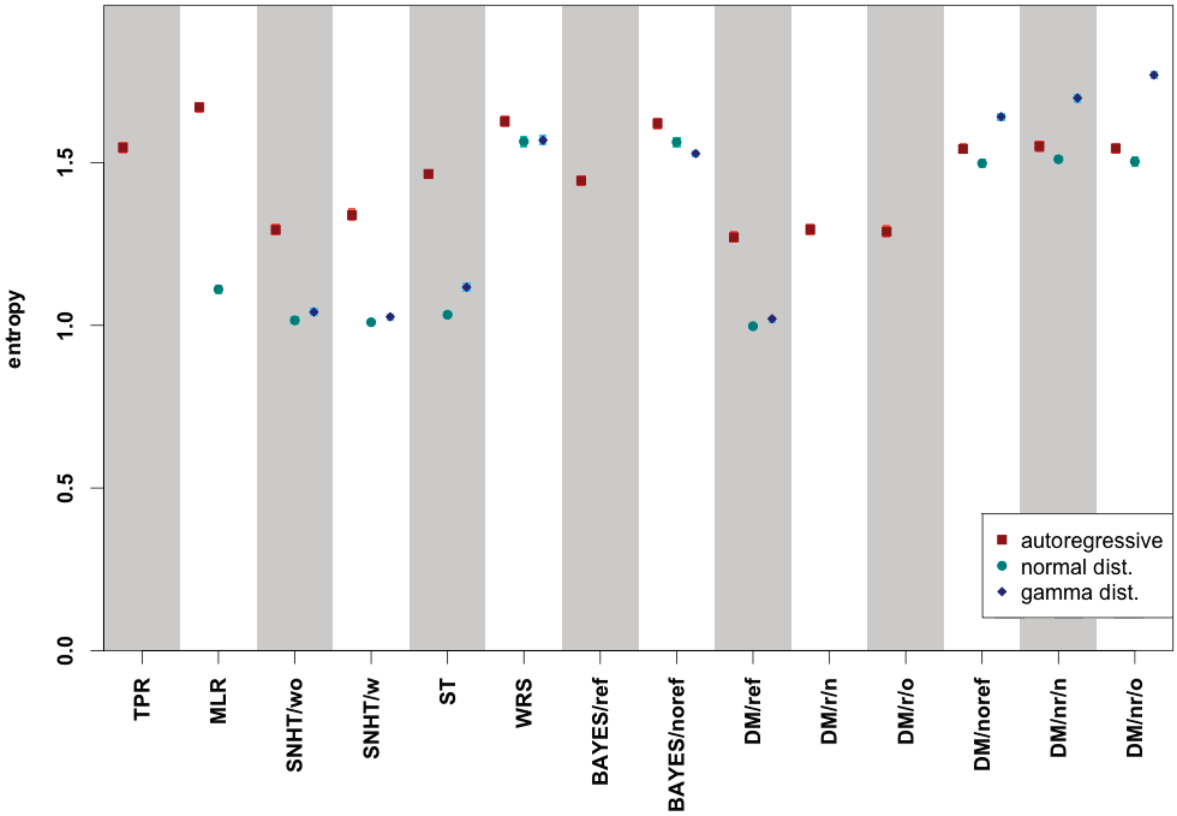


Figure 3.20: Entropy of the inter-comparison of the modifications of the Dose-Menzel method and the methods described by Ducré-Robitaille et al. [2003] on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the entropy. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

lower results deliver the methods that depend on a reference series. Here, especially the results of the normal and the gamma distributed test vectors for the flat model are low. This indicates that a clear decision is made by these methods in these cases. A look at the underlying contingency tables (is shown in the tables A.2 and A.3 in the appendix) shows, that these models deliver only in a few cases a false result. This was also suggested by the results of the log odds ratio and HSS. The other models using a reference series do not show any results for the normal and gamma distributed test vectors. This is due to the fact, that they deliver in one case, here for the inhomogeneous series, a perfect result. Would this be taken into account, the results for these methods for those vectors would be on a comparable level to the ones of the flat model. For the test vectors generated by an autoregressive process, the results for these modifications are at around 1.29. The methods using no reference series on the right hand side of the plot, have higher values for the entropy for all test vectors. This means, that their hit rate is lower and/or their false alarm rate is higher.

Figure 3.20 shows the results of the models described by Ducré-Robitaille et al. [2003] and the main modifications of Dose and Menzel. Here, the previous results are reaffirmed. The only methods, which

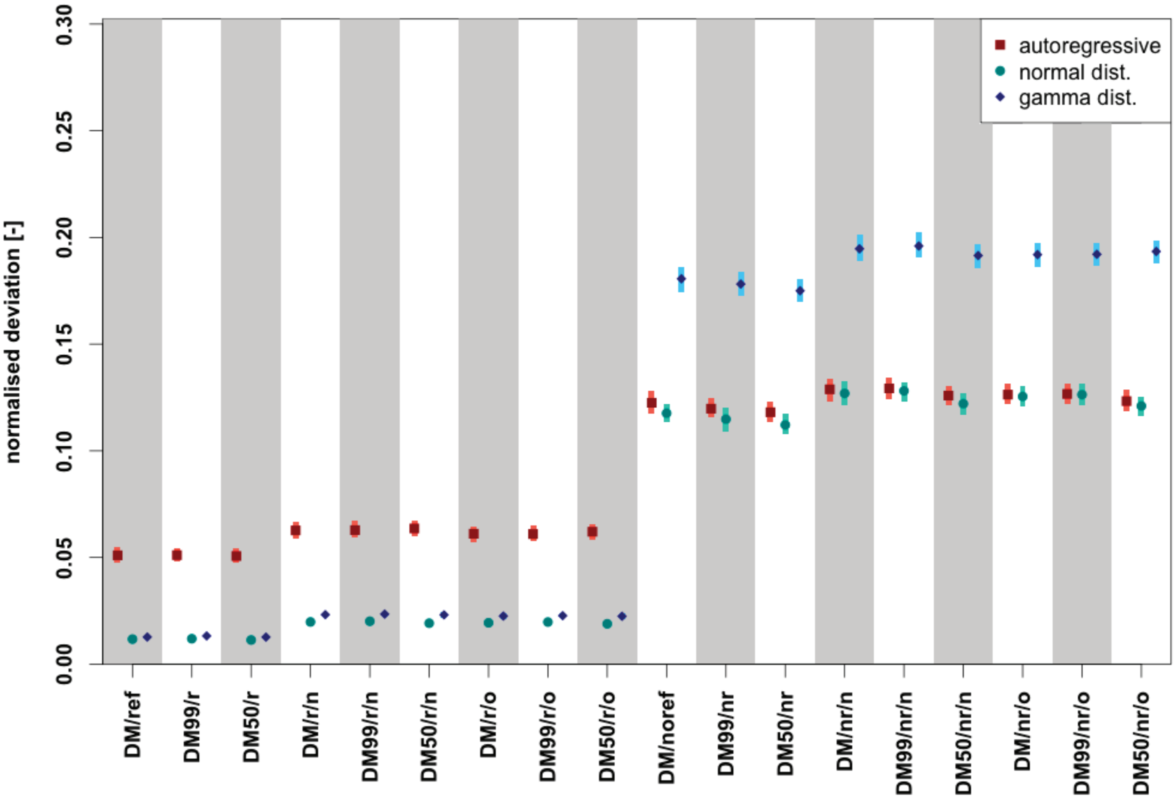


Figure 3.21: Normalised deviation of the position of the steps in the application of the modifications of the Dose-Menzel method on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the normalised deviation. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

deliver similar results to the Dose-Menzel methods with reference series are the two modifications of SNHT. Similar to the methods without reference series, are the results of the WRS and Bayes without reference series. All the other results of the shown methods can be found in between. Remarkable is the fact, that the TPR and MLR have also empty entries in their contingency tables, which was seen for their results of the log odds ratio as well. Both deliver perfect results for the detection of inhomogeneities for the gamma distributed test vectors, while the TPR does the same for the normal distributed.

d) Position of the steps The second part of the analysis will investigate the correct positioning of the steps. Therefore, the detected steps are compared with the true steps, which are included into the time series. To estimate the difference between these two pieces of information they will each be transferred to a cumulative distribution function (cdf). Its creation starts with setting up a vector with the length of the controlled time series. For every position of this vector the number of detected steps up to this position will be included into the vector. By the division of the whole vector by the number of detected steps of the corresponding vector, it is normalised to the cdf.

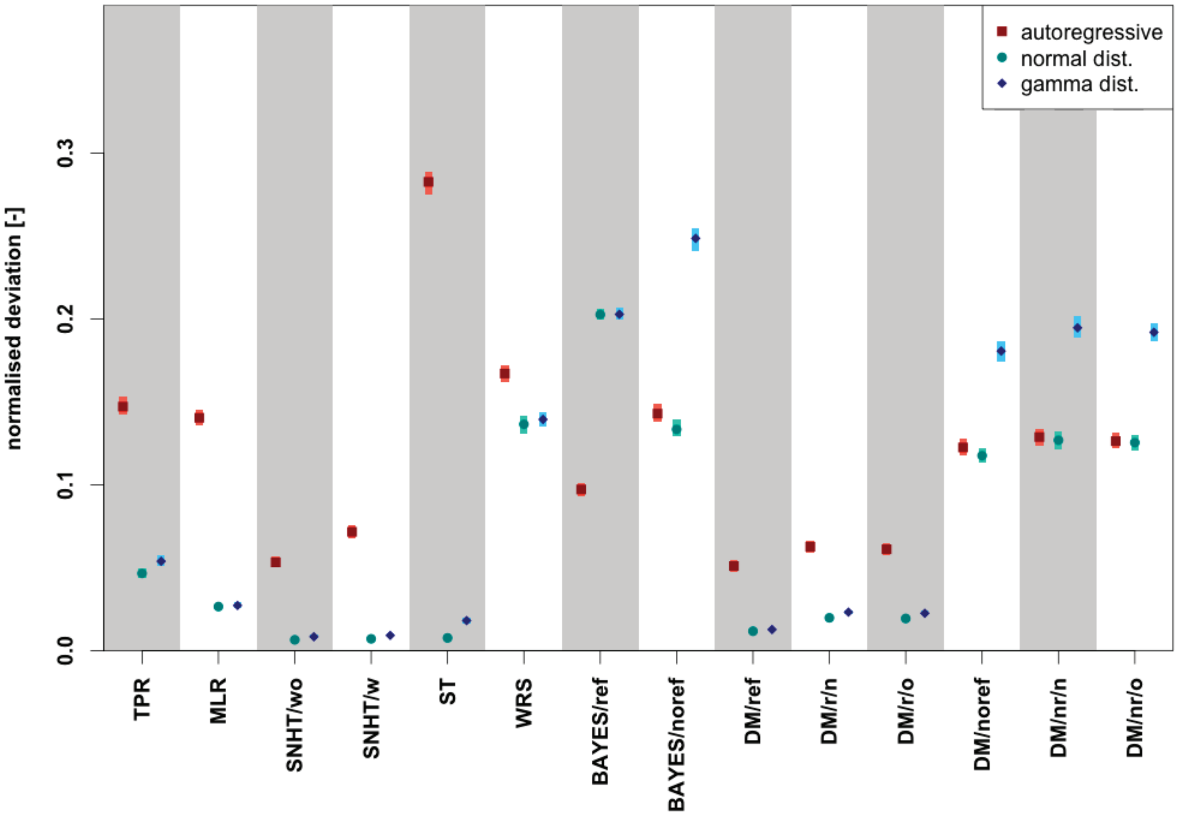


Figure 3.22: Normalised deviation of the position of the steps in the inter-comparison of the modifications of the Dose-Menzel method and the methods described by Ducré-Robitaille et al. [2003] on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the normalised deviation. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

The two resulting cdfs are compared by calculating the equation of the Earth Mover's Distance (equation 3.14). In this case, the number of blocks is given by the number of elements of the vector, which is here $n_b = 100$. Apart from this factor of normalisation, it is the same as the divergence of the Continuous Ranked Probability Score (CRPS) (Gneiting and Raftery [2007]). For only one breakpoint the calculation is equivalent to the standard CRPS (Hersbach [2000]).

The results for the modifications of Dose-Menzel are shown in figure 3.21 and for the comparison of their main modifications with the models described by Ducré-Robitaille et al. [2003] can be found in figure 3.22. The plot has the same layout like the plot before, with the normalised deviation shown on the y-axis. Illustrated are the results for the mean over 25000 realisations. The orientation of the y-axis is chosen such that lower values mean, that the set steps are closer to the original included steps. The normalised deviations for the modifications of Dose and Menzel in figure 3.21 are again divided into the methods using a reference series and those, which do not. For the first, the results of the normal and gamma distributed vectors are much lower than the results for the vectors with the autoregressive process. For

Table 3.6: Properties of Dose and Menzel methods combined with the histogram test used in the sensitivity tests.

Abbrevation	Model	reference series	threshold	histogram test method
DM/ref/KLD	flat	yes	95%	KLD
DM/ref/JSD	flat	yes	95%	JSD
DM/ref/MS	flat	yes	95%	MS
DM/ref/RMS	flat	yes	95%	RMS
DM/ref/EMD	flat	yes	95%	EMD
DM/ref/n/KLD	normal	yes	95%	KLD
DM/ref/n/JSD	normal	yes	95%	JSD
DM/ref/n/MS	normal	yes	95%	MS
DM/ref/n/RMS	normal	yes	95%	RMS
DM/ref/n/EMD	normal	yes	95%	EMD
DM/ref/o/KLD	original	yes	95%	KLD
DM/ref/o/JSD	original	yes	95%	JSD
DM/ref/o/MS	original	yes	95%	MS
DM/ref/o/RMS	original	yes	95%	RMS
DM/ref/o/EMD	original	yes	95%	EMD
DM/noref/KLD	flat	no	95%	KLD
DM/noref/JSD	flat	no	95%	KLD
DM/noref/MS	flat	no	95%	MS
DM/noref/RMS	flat	no	95%	RMS
DM/noref/EMD	flat	no	95%	EMD
DM/nr/n/KLD	normal	no	95%	KLD
DM/nr/n/JSD	normal	no	95%	JSD
DM/nr/n/MS	normal	no	95%	MS
DM/nr/n/RMS	normal	no	95%	RMS
DM/nr/n/EMD	normal	no	95%	EMD
DM/nr/o/KLD	original	no	95%	KLD
DM/nr/o/JSD	original	no	95%	JSD
DM/nr/o/MS	original	no	95%	MS
DM/nr/o/RMS	original	no	95%	RMS
DM/nr/o/EMD	original	no	95%	EMD

the latter, the results are generally much higher, but here, the vectors resulting from the autoregressive process and the normal distributed vectors are lower than for the gamma distributed vectors.

Of more importance are the results of the comparison to the methods described by Ducré-Robitaille et al. [2003], shown in figure 3.22. The main conclusion from this plot is, that the results of the modifications of Dose-Menzel using a reference series are in the same order of the methods with the best performance by Ducré-Robitaille et al. [2003]. This is reached by the SNHT in its two modifications. Also, the ST shows low results for the normal and gamma distributed vectors, but much worse for the autoregressive process. Other methods like TPR and MLR show results with a similar behaviour. WRS and the two Bayes modifications are in the same order like the Dose-Menzel modifications without a reference series.

Further discussion on the results in this section is given in section 5.1.2.4.

3.5 Combination

In the last section, the method of Dose & Menzel was introduced and modified. Some sensitivity tests were performed and at the end, the method was compared to other change point detection methods. As a next step, it will be used in different applications in the next chapter 4. In this section, some preparations for these applications will be made. It will be explained, how to combine the test of Dose & Menzel with the histogram test introduced in section 3.3.

The basic strategy is to extract a one-dimensional time series from the result of a histogram test. The resulting matrix for controlled time series consists of rows and columns for each temporal block, which show the comparisons to the other blocks. For this test, one of these rows or columns is used, that is a one-dimensional time series. As a consequence, from this time series it is possible to determine, how the chosen block compares to all other blocks. Resulting from that, it becomes possible to find inconsistencies within the dataset. For observational data it is a good choice to take the last existing row or column, which might indicate the latest observation. In this case, it is normally best known how the involving instruments behave and which problems are introduced into the measurement process. For simplicity, only the newest row is used in the following tests. The difference between choosing a row or column would only be a problem for asymmetric measures like the KLD, what was explained in section 3.3.3.1. In this section, only shifts in mean will be analysed. In section 3.3.3.3 it was described, that the difference with the KLD for such a shift is not high enough within the results, that this would be a problem for the following comparison tests. Therefore, the KLD can be used like the other distance measures, without the introduction of a bias by taking the wrong choice on which parts of the result matrix the evaluation take place.

The two tests, that will be performed again, are the homogeneity test from section 3.4.4.3 and the step detection test from section 3.4.4.4. The used test vectors have a length of 100 elements. For the histogram test, this is a very short time series, as it divides the series into blocks and estimates with a histogram for each a probability density function. To detect inconsistencies here, a size of block $s_b = 5$ is used. This means, that the pdf has to be estimated from only five elements. The other parameters are chosen like before, the number of bins is $n_b = 65$ and the prior information used for the KLD and JSD is $a_f = 100$. The extracted time series is the last available row of the result matrix. In this case the last element of the time series is a comparison of the block with itself. This was excluded in the analysis and therefore the time series has a length of 19 blocks. It is then analysed with the modifications of Dose & Menzel. In the following tests, all combinations of used measures in the histogram tests and basic modifications of Dose & Menzel are calculated and compared.

3.5.1 Test on homogeneous datasets

The procedure for this test was described in section 3.4.4.3. In the same section, the structure of figure 3.23, which shows the results of this test, was explained. The x-axis shows the different methods, the y-axis the percentage of inhomogeneities. Used abbreviations for the here introduced methods are described in table 3.6. Abbreviations for the basic modifications of Dose & Menzel were already included in table 3.4. Methods that use a reference series subtract the reference series from the original series before the histogram test is applied.

The plot shows the thirty combined methods alongside the six basic modifications. At first sight, mainly the methods using the Kullback-Leibler Divergence deliver a high number of inconsistent time series for the homogeneous time series. Especially for the normal model, where intercept and slope is regressed within the change point detection of Dose and Menzel, the results are high. The other methods do not deliver inhomogeneities for the flat model with and without reference series. This is different to the

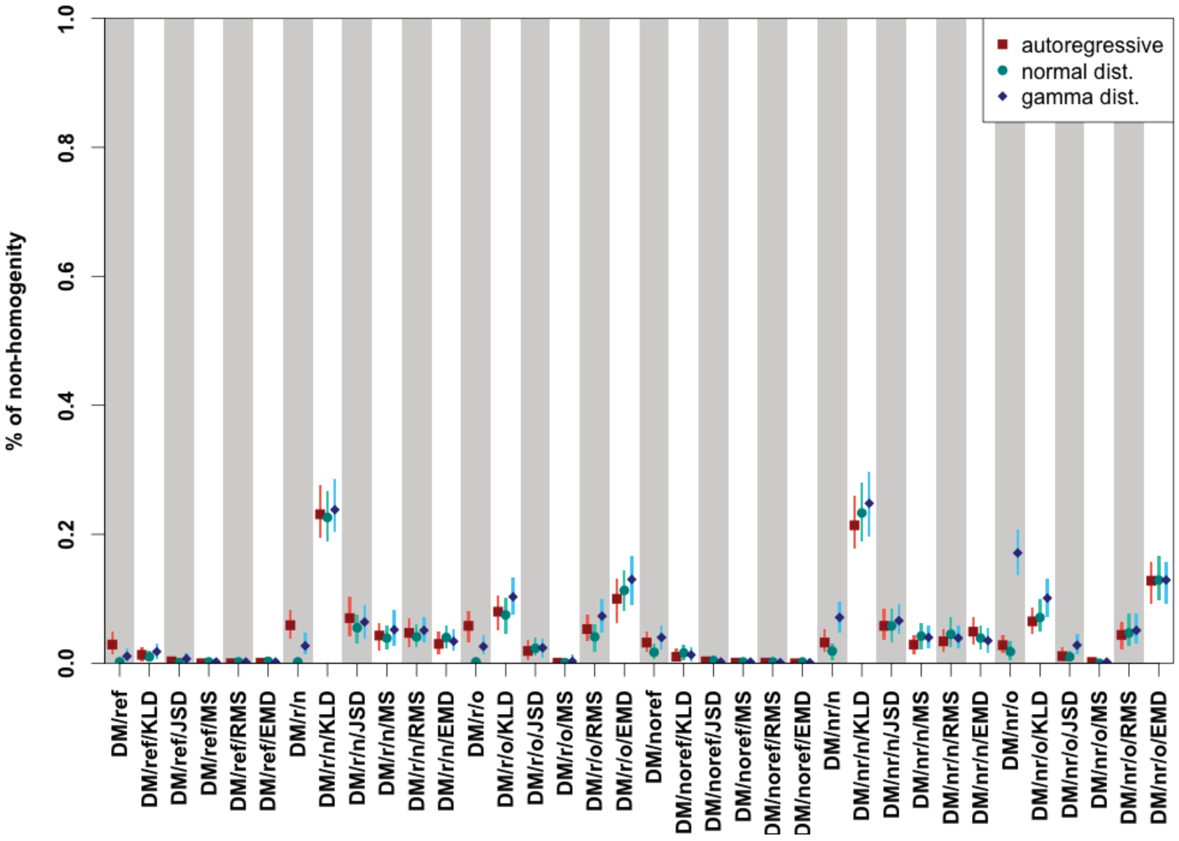


Figure 3.23: Testing homogeneous datasets with the modifications of the Dose-Menzel method on results of the histogram test. On the x-axis the methods are shown, on the y-axis the percentage of detected inhomogeneities. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the value over 1000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

case, where the Dose & Menzel method is applied to the same series without using the histogram test. There, the methods reach between 0.2 and 4.0%, respectively. For the normal model, again only the KLD method deliver higher results than the original modification. The others are more or less in the same order as the autoregressive process for the basic modification. Remarkable is the absence of big differences between the three types of test vectors for the methods using a histogram test in this case. For the original methods by Dose & Menzel the results are very heterogeneous. Here, mainly the KLD and EMD deliver high results of around 7.5 to 13 %. The others are mainly in the same range like the original modifications. Here, also the differences between the different types of test vectors are low for a given method.

Summarising the results of this figure leads to the following conclusion. Except the KLD, the main measures deliver similar numbers of as inconsistent classified time series like the original modification. Remarkable is, that the difference between the different types of test vectors is lower than for the original modification without using the histogram test.

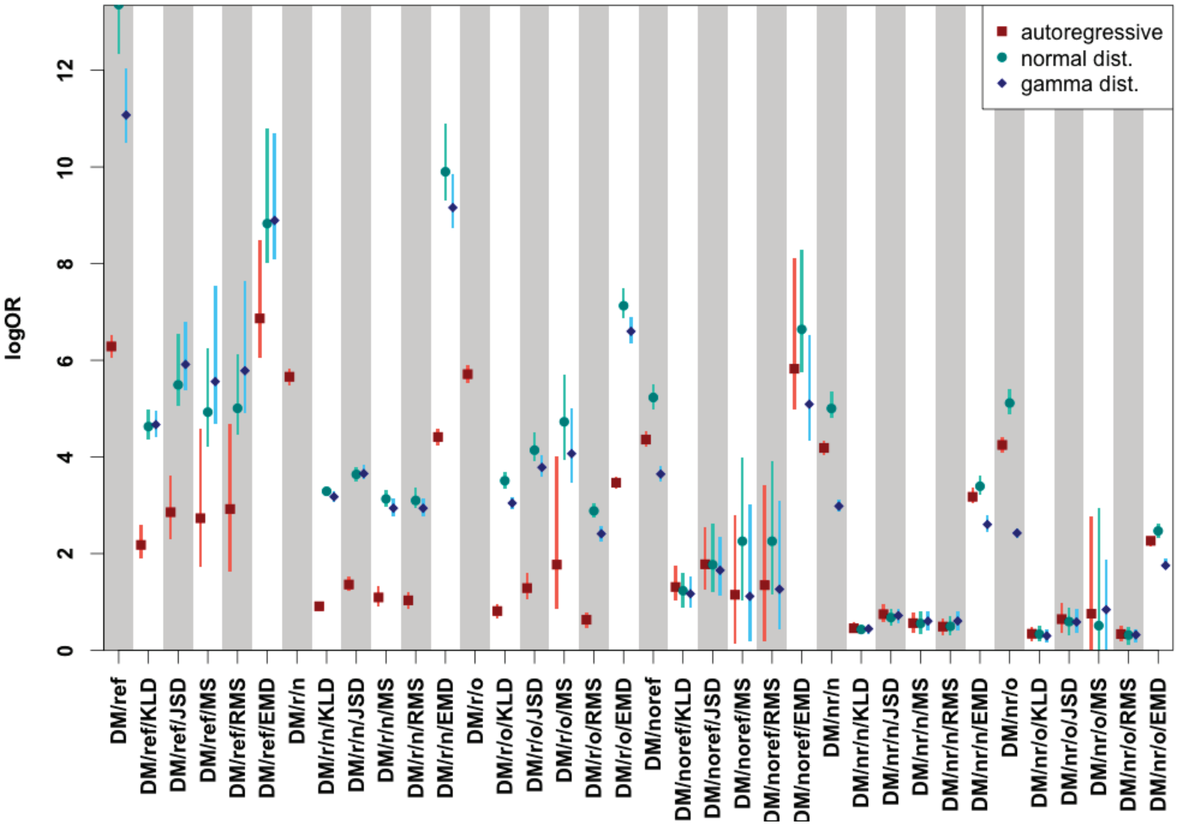


Figure 3.24: Log odds ratio of the application of the modifications of the Dose-Menzel method in combination with the results of the histogram test on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the log odds ratio. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

3.5.2 Detection of change points

The results for the test on inhomogeneous data is shown in the figures 3.24 to 3.26 and include again the results for the log odds ratio, Heidke Skill Score and entropy.

a) Log odds ratio Figure 3.24 shows the results for the log odds ratio. The plot has the same structure like the last shown figure. On a first glance, the outcome for the different methods seem to be very heterogeneous and complicate. Therefore, the results for each of the six basic modifications will be discussed separately.

It should be started with the flat model, which uses a reference series and is located on the left hand side of the plot. Obvious are the high uncertainties of all methods using the histogram test. For the autoregressive test vectors, only the EMD measure is in an order of the original method. All others are only around half as high. Also for the normal and gamma distributed vectors, the results are higher for the EMD than for the ones performed with the histogram test. The difference between those two

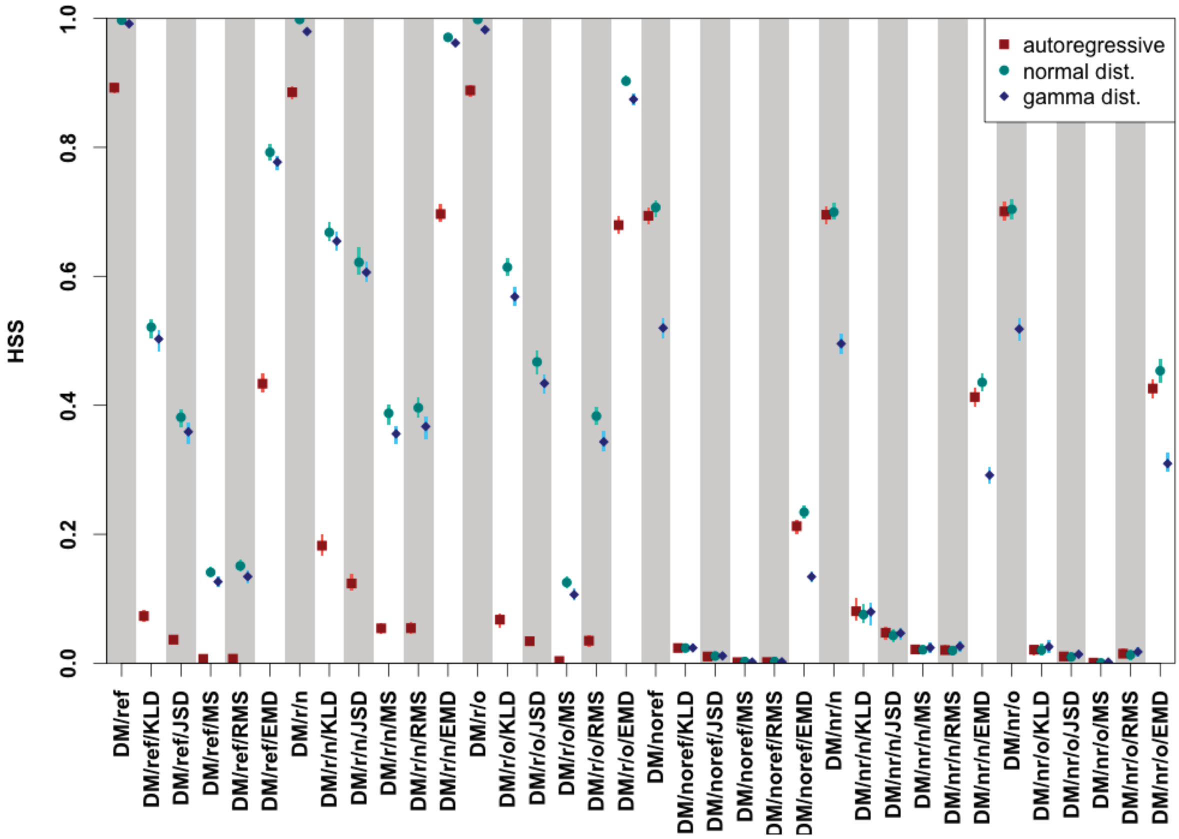


Figure 3.25: Heidke Skill Score of the application of the modifications of the Dose-Menzel method in combination with the results of the histogram test on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the Heidke Skill Score. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

results is for all histogram methods much lower than for the original method. For the normal method with reference, the uncertainties are much lower. In this case, the EMD is the only method, that delivers results comparable to the original modification. The other methods using the histogram test deliver results for the autoregressive process of just around 1. In case of the third modification, the original model with reference, the result of the autoregressive process is in all cases of the histogram test clearly lower than the results of the version of the Dose & Menzel test, which is applied to the original vector. The best is again the EMD measure. The MS measure shows a large uncertainty. In cases of the normal and gamma distributed vector, where the original modification has not delivered any results, the EMD performs best as well.

The methods, which do not use the reference series, deliver generally similar results to the ones using it. For the flat model the uncertainties are also high for the methods, using the histogram test. In this case, the EMD version delivers higher results for all types of test vectors than all other modifications, including the original modification. For the normal model, the results are much more heterogeneous without a reference series than with it. Here, the results for all types of test vectors are for all methods

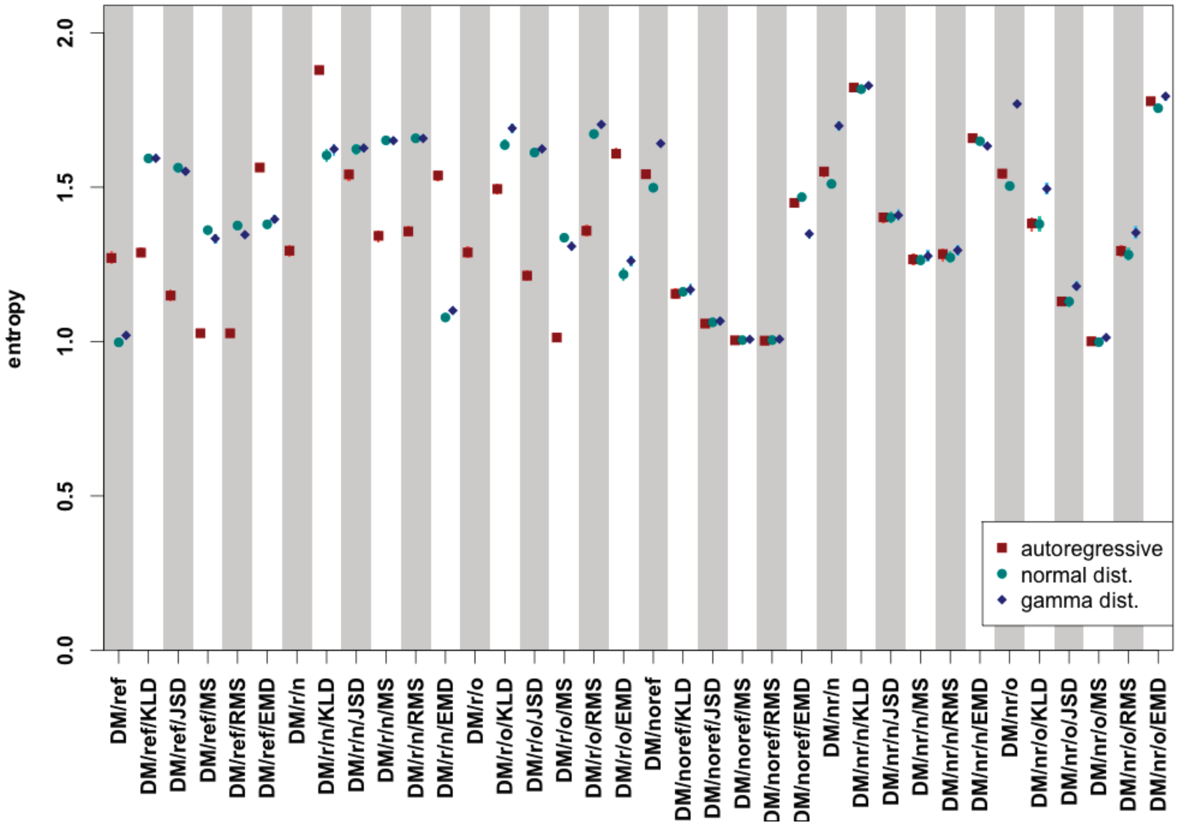


Figure 3.26: Entropy of the application of the modifications of the Dose-Menzel method in combination with the results of the histogram test on homogeneous and inhomogeneous test vectors. On the x-axis the methods are shown, on the y-axis the entropy. Three different types of test vectors are used: autoregressive (red), normal distributed (green) and gamma distributed (blue). The marks indicate the mean of 25000 test vectors, the bar behind it the uncertainty estimated by bootstrapping with 1000 samples.

very similar. The EMD delivers slightly lower results than the original modification. The same holds for the original model with or without the use of a reference series.

b) Heidke Skill Score The results for the Heidke Skill Score in figure 3.25 show a similar behaviour for the different methods. It can be seen, that especially for the normal and the gamma distributed test vectors, the normal method of the EMD measure with the use of a reference series deliver high results (0.971 and 0.962). Together with its equivalent for the original method this method delivers results similar to the original methods without reference series for the autoregressive test vectors (DM/r/n/EMD: 0.697, DM/r/o/EMD: 0.679, DM/noref: 0.694, DM/nr/n: 0.696, DM/noref/o: 0.701). For most of the methods, the results are better for the normal and gamma distributed vectors than for the autoregressive process. Also the difference between those two test vector types is relatively low for most methods using the histogram test.

c) Entropy For the entropy the results in figure 3.26 are very heterogeneous. The EMD, which performed so well in the two figures before, now show values higher than 1.5 for the most methods of Dose and Menzel. This indicates, that the methods do not deliver clear results. Reasons can be found, when the contingency tables are further analysed (table A.1 to A.3 in the appendix). They show, that the problems mainly occur for the inhomogeneous datasets. For the homogeneous datasets the results are clear, what can be interpreted as a low false alarm rate on inhomogeneities. The other measures of the histogram tests partly show a much lower entropy, but this does not necessarily have to be an indication for a good result. It might also be possible, that a high rate of false alarm or misses can be found within the results. A look at the results for the contingency tables in table A.1 to A.3 in the appendix underline this interpretation for most of the histogram methods, which are not combined with the EMD.

Summarising the results of all plots leads to the conclusion, that the use of the histogram test with the EMD leads to relatively small changes to the results compared to the original method. On the other hand, it is not as sensitive to inhomogeneities in datasets as the original method. The other measures deliver a much lower performance in terms of the log odds ratio and HSS. The entropy shows, that the methods have a different detection limit than the original method. This requires a recalibration of the method with the histogram test, in order to obtain better results. Nevertheless, it has to be kept in mind, that the histogram test in general performs a strong data reduction. To get a precise position and a verification of the potential steps, the breaks have to be checked again with the original method, but this time without the combination with the histogram test. The estimation of the probability density functions with the histograms of only five elements is also critical.
More discussion on that topic follows in section 5.1.2.4.

4 Application

In chapter 3 several checks for a quality assurance on data were presented. Additionally, some tests on artificial datasets were performed, making use of those checks. In this chapter, the checks will be applied to different datasets of real observations and model data. The aim is to demonstrate, how these tests can be applied in real case situations and which further enhancements can modify them to make the tests usable in new fields of applications. In section 4.1, the histogram test in combination with the Dose and Menzel methods is applied to observational data. Two different measurements of wind will show that this approach is able to detect uncertainties due to rounding effects within a dataset. Section 4.2 shows the analysis of inconsistencies in multidimensional datasets with the help of the combined method. In this analysis, the surface windspeed from NCEP and ERA40 reanalyses dataset is taken as an example. In a third application in section 4.3, the combined method is used to demonstrate a parallel analysis of a large number of datasets. It is performed by the analysis of the basic data for the HADCRUT 3 dataset. Additionally, it is shown how the resulting temperature reconstruction handles inconsistencies in the initial datasets. In a last application, the quality evaluation is demonstrated on data of a climate station in section 4.4. It uses the methods of Meek and Hatfield [1994] and combines their results with assumptions on how these methods give information about the quality of the datasets.

4.1 Detection of rounding in data

The first application shown in this chapter is the application of the histogram test on a meteorological time series, which was generated by a meteorological station operated by the German Weather National Service (DWD) in Lindenberg (Germany) (station id: 10393 / latitude: 52° 21' North / longitude: 14° 12' East , elevation: 98 m). The test is performed by the evaluation of the mean wind and daily maximum wind from this station. Both time series have been measured over twenty years, between 1991 and 2010. They are shown in figure 4.1.

For the application of the histogram test, the different measures described in section 3.3.2 are used. The block length is chosen as $s_b = 365$ values and the number of bins for the histograms as $n_b = 65$. For the Kullback-Leibler divergence and the Jensen-Shannon divergence the prior a_f is set to 100. The motivation for the choice of the parameter s_b with a length of one year is to prevent possible problems arising due to the annual cycle. The results for four of the five measures (KLD, JSD, RMS, EMD) are shown in the subfigures 4.2a-d and 4.3a-d. A subsample of the datasets with a focus on detected inconsistencies is shown in the subfigures 4.2e and 4.3e.

The bin-wise comparing measures, KLD (subfigure 4.2a), JSD (4.2b) and RMS (4.2c) for the mean wind in figure 4.2 show a pattern in the result matrix. This indicates that a change around the years 2000/2001 has occurred in the dataset. In the plot for the results of the EMD (subfigure 4.2d) are hardly any indications for a break at this time recognisable. To show what leads to the inconsistencies in the results of the three measures, a focused extract on the dataset for the time span between 2000 and 2002 is shown in subfigure 4.2e. Obviously, a change in the storing of the dataset happens at 1st April, 2001. This change is indicated by the rounding of the data after this datum, which was not applied before.

In the following step, the combination of the the histogram test and the methods of Dose and Menzel

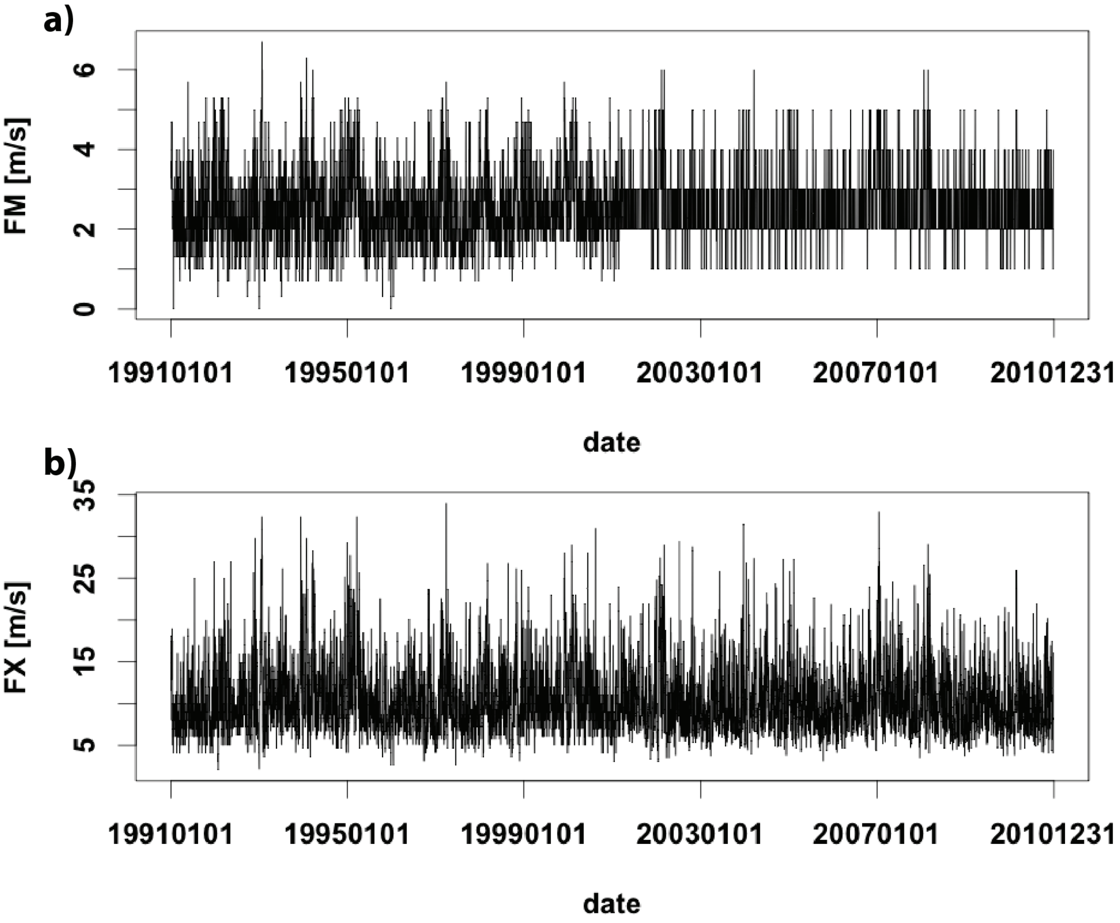


Figure 4.1: Original time series of the wind speed measurements in Lindenberg between 1991 and 2010. In the upper subfigure the mean wind and in the lower the daily maximum wind are shown.

will be demonstrated. The results for each of the five measures and regression models within the Dose and Menzel method are shown in table 4.1. For each combination, the probability of a change point, the position of the most probable breakpoint and the probability of a break at this breakpoint are shown. The models used for the Dose and Menzel method are the calibrated methods with a 95% threshold. In case that the probability of a change point is lower than 95%, the most probable year and the probability of a change point in the corresponding year are given in brackets.

All measures, except the EMD method, show a clear indication of a break for all types of regression models in the year 2001. For the EMD, all regression models deliver a probability for a change point, which is lower than 40%. Nevertheless, the flat model in the year 2001 shows with 62% a small advantage of the change point model compared to the constant model.

As a second example, the daily maximum wind for the same station is investigated. The results are shown in figure 4.3. Again, the bin-wise working methods (KLD, JSD and RMS) show certain patterns. These indicate a difference within the data of the years 1991, 1999 and 2000 to the other years covered by the datasets. In contrast, the visual inspection of EMD does not show any obvious pattern. The reason for these patterns can be seen within subfigure 4.3e, which shows a focus on the time span between July 1998 and July 2001 of the dataset. It also shows a rounding, which can be recognised between December 1998 and April 2001. Before and after this time span, the data are stored differently. For similar reasons

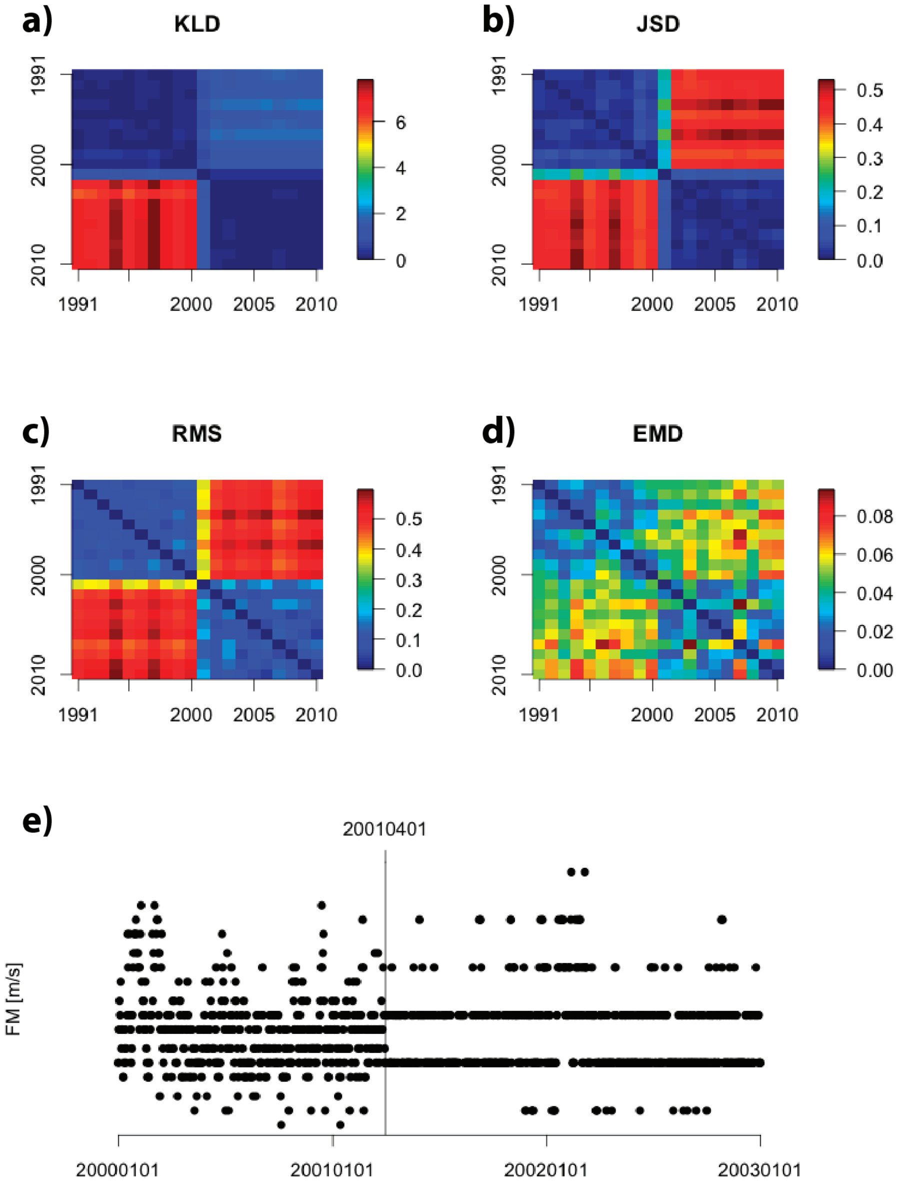


Figure 4.2: Application of the histogram test with the different measures on the mean wind measured in Lindenberg from 1991 to 2010. In the four upper figures, the result matrix for the KLD (upper left), JSD (upper right), RMS (middle left) and EMD (middle right) are shown. In the figure at the bottom, the time frame between January 2000 and December 2002 of the original values is shown.

Table 4.1: Analysis of the mean wind measured by the meteorological station in Lindenberg. Illustrated are the results of the application of the combination of the histogram test and the different regression models within the change point analysis by Dose and Menzel.

model	information	KLD	JSD	RMS	MS	EMD
flat model	probability of change point	1.000	1.000	1.000	1.000	0.148
	best candidates	2001	2001	2001	2001	(2001)
	probability of change at year	1.000	1.000	1.000	1.000	(0.619)
normal model	probability of change point	1.000	1.000	1.000	1.000	0.375
	best candidates	2001	2001	2001	2001	(2001)
	probability of change at year	1.000	1.000	1.000	1.000	(0.294)
normal model	probability of change point	1.000	1.000	1.000	1.000	0.022
	best candidates	2001	2001	2001	2001	(2001)
	probability of change at year	1.000	1.000	1.000	1.000	(0.175)

the patterns in the results of the three measures in 1991 are generated.

In table 4.2, the results of the application of the Dose and Menzel modifications on the data of the year 2010 are shown. Obviously none of the methods is able to detect a change point within the dataset. With around 70% probability for that a change point exists at all, the clearest indication for all types of regression models is given by the KLD for the year 2000/2001. In the specific year, the probability is given by a maximum of 94%. Since the KLD measure is an asymmetric measure, like in the sensitivity tests in section 3.3.3, the best result of two possibilities is taken. All other measures are far below these results. Most methods combined with the different measures show the highest probability of a change point in the time span between 2000 and 2002. An exception is the EMD which shows 1994 as the best candidate. Nevertheless, it delivers a very low probability for all modifications of Dose and Menzel for this year.

Summarising the results of these applications leads to the conclusion, that for the detection of rounding in data the bin-wise methods are superior to the EMD. For obvious patterns, the combination of the histogram test with the change point detection by Dose and Menzel delivers results, which are comparable to the visual inspection. Nevertheless, the method might have problems with more complicate patterns. Further discussion on the reasons of this behaviour will take place in section 5.1.1.2.

4.2 Inconsistency detection of reanalysis data

This section shows a second application of the histogram test. In this case, the time series under consideration do not have one, but three dimensions. The datasets, to which the methods are applied, are reanalysis data of NCEP in section 4.2.1 and ERA40 in section 4.2.2. Investigated is the surface wind representation within these model data.

The surface wind speed parameter in the two global reanalyses have already been investigated by several studies. An example is the analysis of Monahan [2006], which was performed with the help of probability density estimations. Monahan calculated the spatial distribution of different statistical moments and compared them between the different reanalyses. In another study Yuan [2004] indicated, that especially

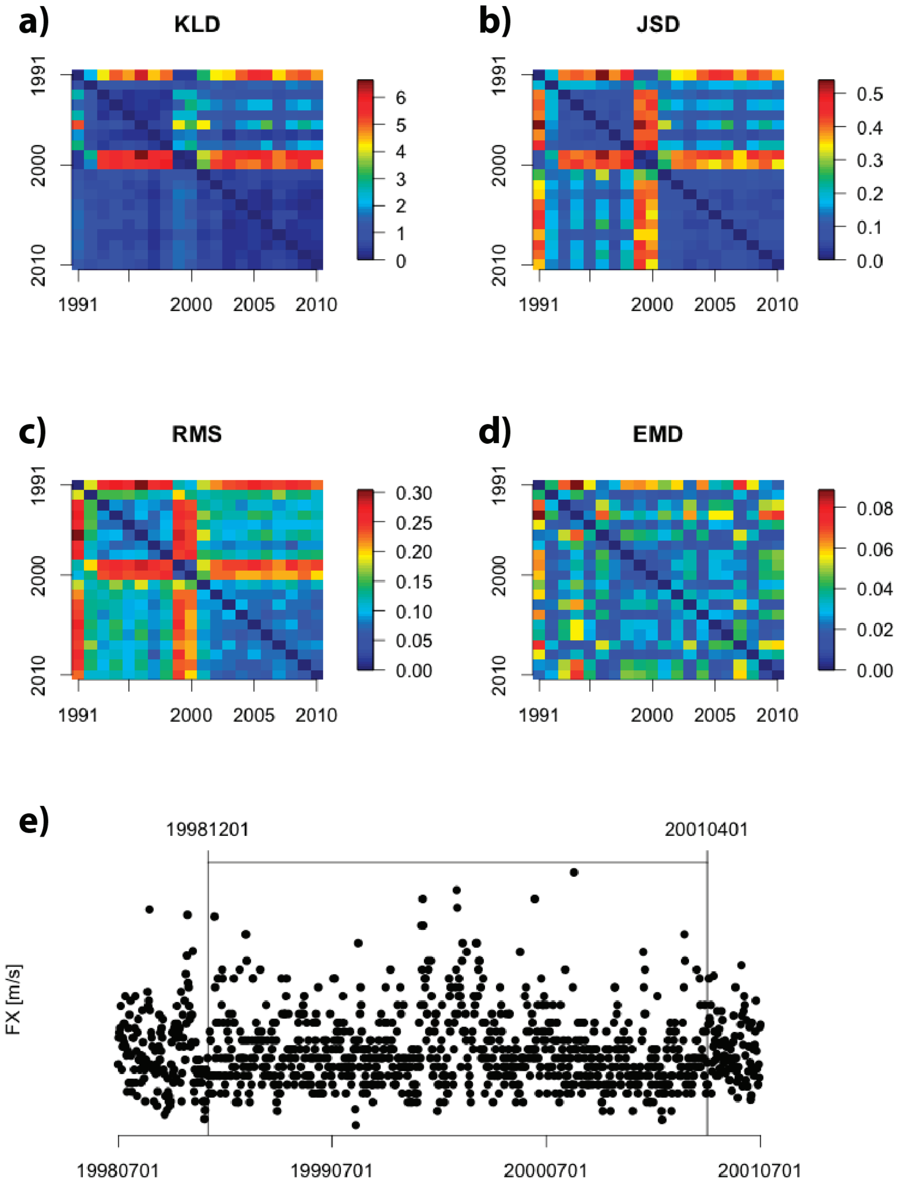


Figure 4.3: Application of the histogram test with the different measures on the maximum wind measured in Lindenberg from 1991 to 2010. In the four upper figures, the result matrix for the KLD (upper left), JSD (upper right), RMS (middle left) and EMD (middle right) is shown. The figure at the bottom illustrates the time frame between July 1998 and June 2001 of the original values.

Table 4.2: Analysis of the maximum wind measured by the meteorological station in Lindenberg. Illustrated are the results of the application of the combination of the histogram test and the different regression models within the change point analysis by Dose and Menzel.

model	information	KLD	JSD	RMS	MS	EMD
flat model	probability of change point	0.693	0.147	0.151	0.011	0.003
	best candidates	(2001)	(2001)	(2002)	(2001)	(1994)
	probability of change at year	(0.936)	(0.476)	(0.501)	(0.042)	(0.004)
normal model	probability of change point	0.633	0.150	0.132	0.012	0.003
	best candidates	(2000)	(1997)	(1997)	(1997)	(1994)
	probability of change at year	(0.888)	(0.456)	(0.384)	(0.046)	(0.018)
normal model	probability of change point	0.749	0.269	0.270	0.024	0.001
	best candidates	(2000)	(2000)	(2000)	(2001)	(1994)
	probability of change at year	(0.939)	(0.738)	(0.798)	(0.019)	(0.011)

in the southern oceans the discrepancies between observations and current reanalyses are high. Yuan emphasised, that this finding depends on the season and is also detectable in the monthly mean data, where the reanalyses underestimate the strength of the winds.

4.2.1 NCEP reanalysis

The NCEP reanalysis was generated by the National Centers for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR) in the United States. Initially, it covered the years 1957 to 1996, but was extended afterwards. The system collects several types of observational data and after a quality control and a data assimilation, they will be processed in a numerical weather prediction model (Kalnay et al. [1996]).

The data is stored as a three dimensional field for each variable of the reanalysis. These fields consists of two dimensions in space and one in time. The chosen temporal resolution of the data is monthly. As a check for inconsistencies the histogram test is used, but unlike the first application it is only applied with the EMD measure. For the comparisons, the data have to be divided into blocks. The chosen resolution in the temporal dimension of these blocks is one year. As a consequence, all data within one year consisting of twelve two-dimensional fields, are used as one block. A weighting is not applied to the data. The blocks are used to calculate the histograms and to compare each block to the other blocks afterwards. The result matrix for this analysis is shown in figure 4.4a.

On the x- and y-axis, the years of the blocks are shown. The values indicate blue for low and red for high differences between the histograms, measured by the EMD. At first glance, some patterns can be recognised. The first is obviously on the upper left hand corner, between the years 1948 and 1957. Others can be found along the diagonal and are partly more and partly less visible. Clear patterns are also located in the years between 1976 and 1978, 1986 and 1995 and after 1996. For further analysis, the dataset can be divided into several subparts. Therefore, the monthly two-dimensional fields are divided into five sections on the meridional direction and the method is separately reapplied to the data of each section. These sections are the tropics (25°S - 25°N , subfigure 4.4b), northern mid-latitudes (25°N - 65°N, 4.4c), southern mid-latitudes (25°S - 65°S, 4.4d) and the polar regions in the northern (65°N -

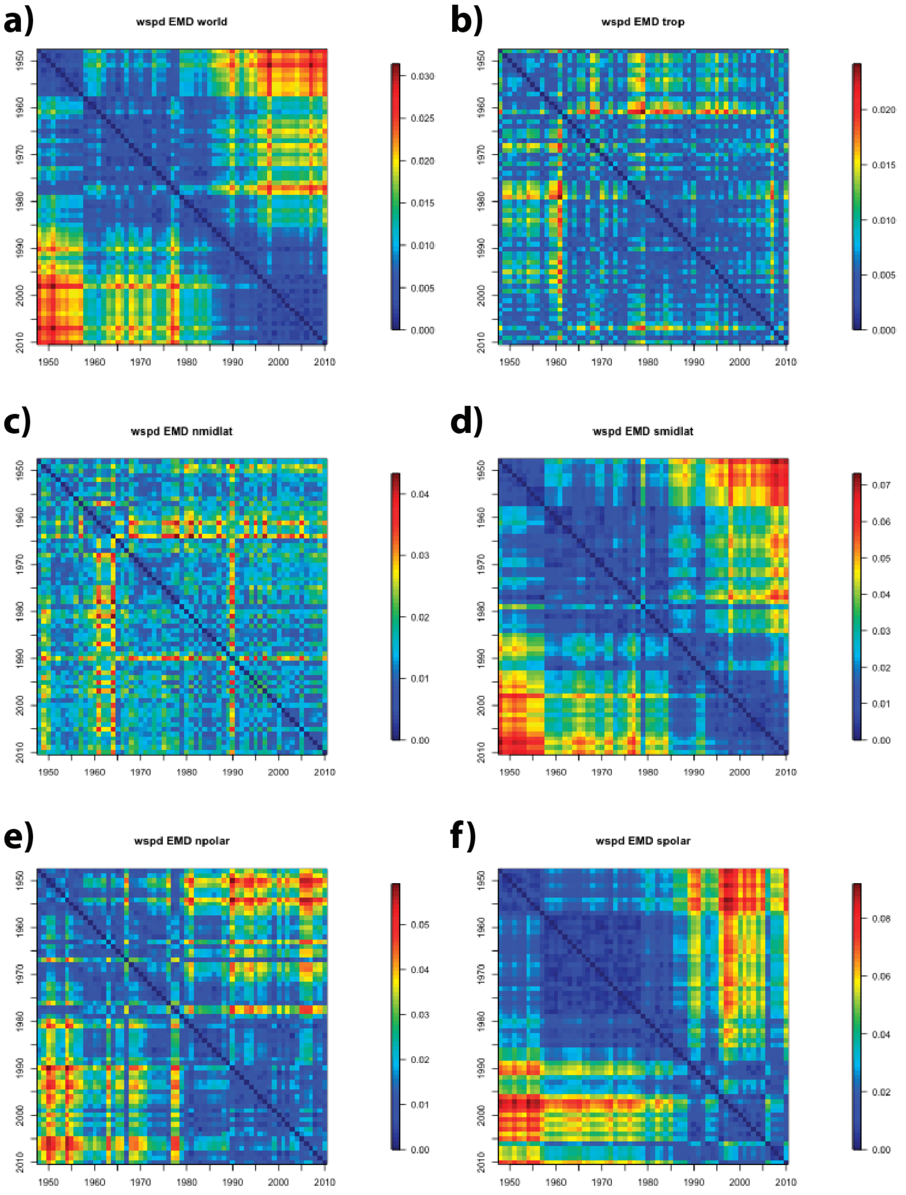


Figure 4.4: Application of the histogram test with the EMD measure to different regions of the NCEP reanalysis surface wind speed for the years 1948 to 2010. In the upper left figure, the result matrix for the whole dataset is shown. Furthermore, the results for the tropical region (upper right), northern latitudes (middle left), southern latitudes (middle right), northern polar region (lower left) and southern polar region (lower right) are presented.

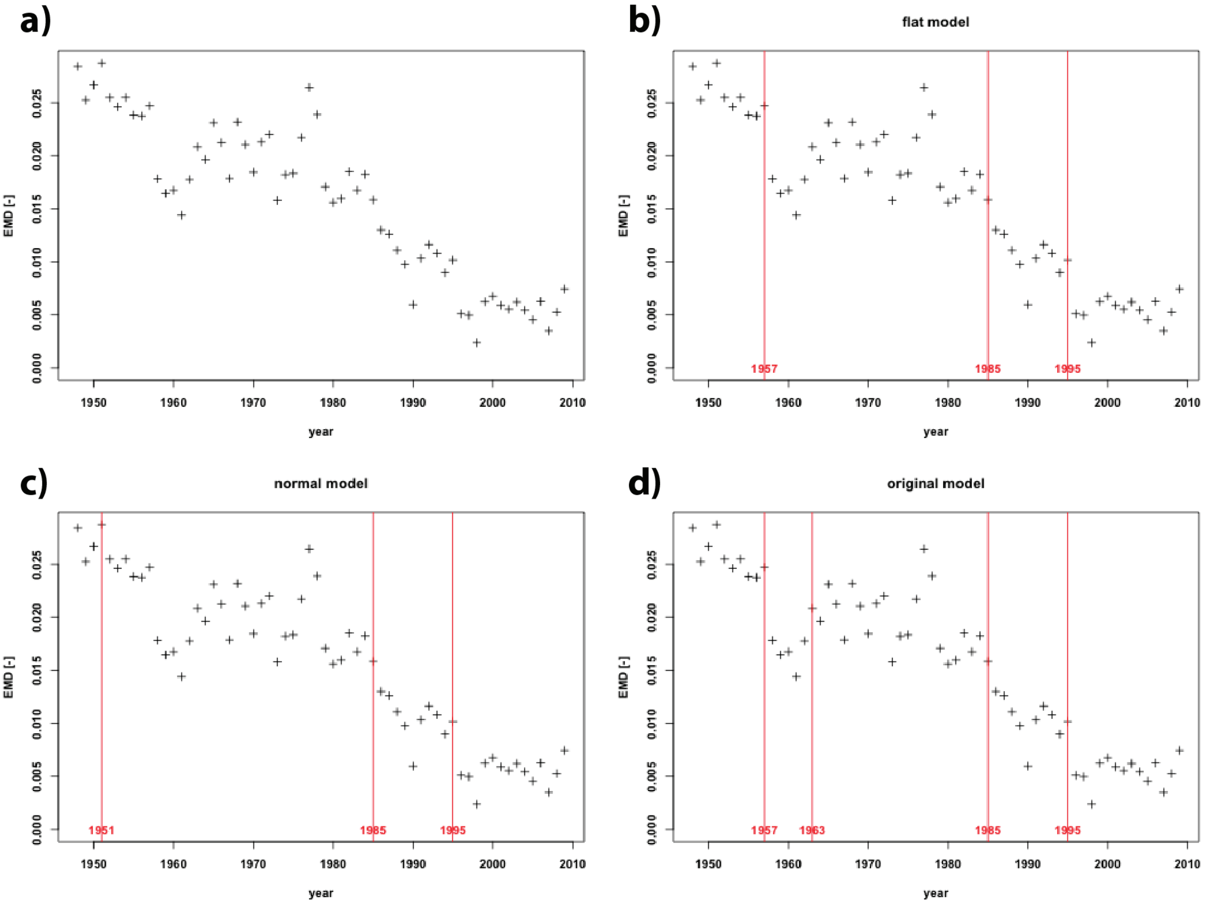


Figure 4.5: Analysis of the results of the histogram test with the EMD measure in comparison to the year 2010 on the global NCEP reanalysis surface wind speed and application of the different modifications of Dose and Menzel. In the upper left, the raw time series is shown. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right, the change points of the original model are shown.

90°N, 4.4e) and southern (65°S - 90°S, 4.4e) hemisphere. All matrices illustrate different results and patterns, so that the inconsistencies of the whole dataset can be attributed to the different regions.

In the tropics some minor patterns are recognisable. The most apparent can be found between 1976 and 1979. The northern mid-latitudes do hardly have any patterns. At some points, the comparison of consecutive years delivers low values, but no larger section is clearly recognisable. For the southern mid-latitudes the results are different. In the year 1957, a strong break occurs in the dataset. Afterwards, several smaller and larger patterns are apparent. Additionally, the year 1979 falls out of its environment of a longer time span of consistent data between 1958 and 1984. The polar region in the northern hemisphere is much more complex. Several minor patterns are recognisable. Additionally, the years 1967, 1978 and 1979 stand out in their environment. In the last plot for the southern polar region, the patterns are structured clearer. Breaks can be found in 1957, 1985, 1988, 1992, 1995, 2005 and 2009.

Before the reasons for these inconsistencies will be explained, the combination of the histogram test and the method of Dose and Menzel is used to analyse the global dataset. The change point detection method is applied to the time series resulting from the histogram test, which includes the comparisons to the data of the year 2010. The corresponding time series is shown in figure 4.5a. The plot has the years on the

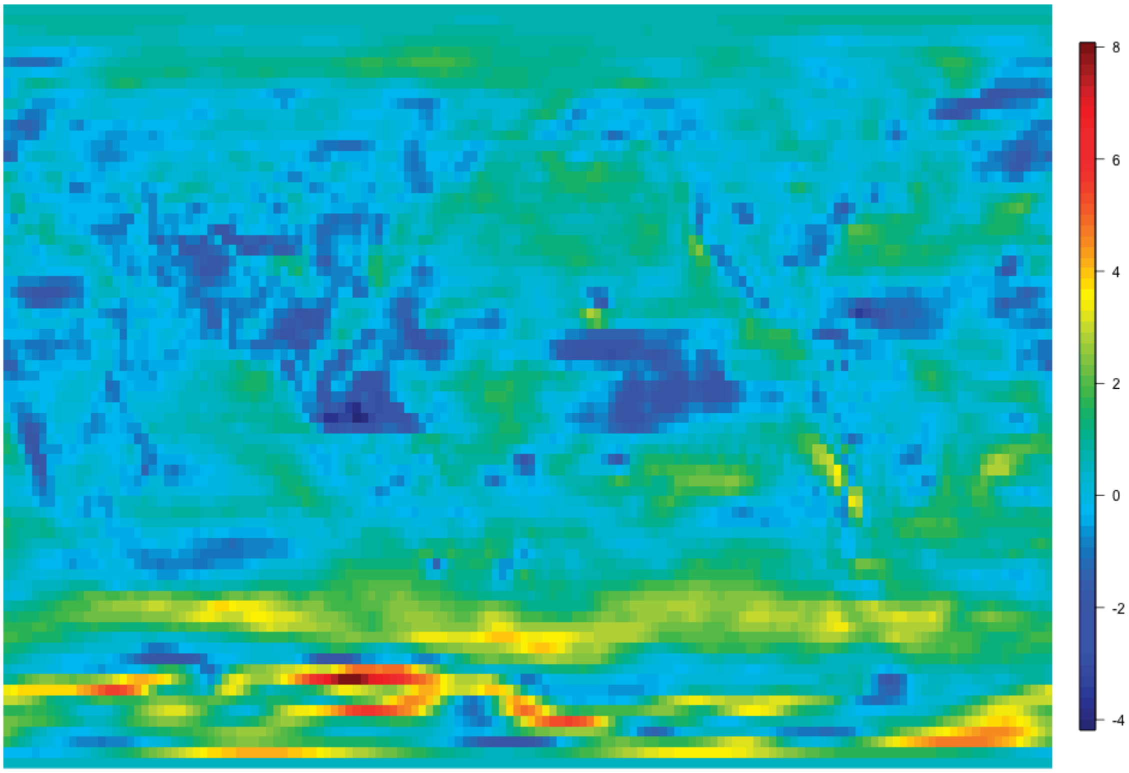


Figure 4.6: Global difference of the yearly mean wind speed between the years 1950 and 2010, measured in metres per second. Positive values in red indicate higher values for windspeed in 2010, negative values in blue indicate higher values in 1950.

x-axis and the values of EMD of the comparisons to the data of 2010 on the y-axis. In general, a trend from higher values in the 1950s to lower values in the recent years can be recognised. The application of the Dose and Menzel methods with the different regression models delivers three to four breakpoints in the time series. The flat model in figure 4.5b set their three change points to 1957, 1985 and 1995. The normal model (figure 4.5c), that also indicates three change points, has the same except for the first, which is set to 1951. The original model, shown figure 4.5d, has the same breakpoints as the flat model, but adds a further one to the year 1963.

The question arises, if the detected breakpoints can be traced back to the generation of the dataset. The first break, which attracted the attention both by visual inspection and the flat and original model, was the break in the year 1957. This year was, together with its successor, the International Geophysical Year (IGY). In this year "[...]many Antarctic stations began collecting data[...]" (Kalnay et al. [1996]). The time before was described by Kistler et al. [2001] as: "The pre-1958 era is [...] the least reliable period, especially for the [Southern Hemispere], where high correlations between analysis and forecasts in 1948 are simply the result of a lack of observations, that is, the 'reanalysis' is mostly a model forecast." The time corresponding to the second feature between 1976 and 1978, which was not indicated by the models, but visible in the result matrix, can be linked to the First GARP Global Experiment (FGGE), which was

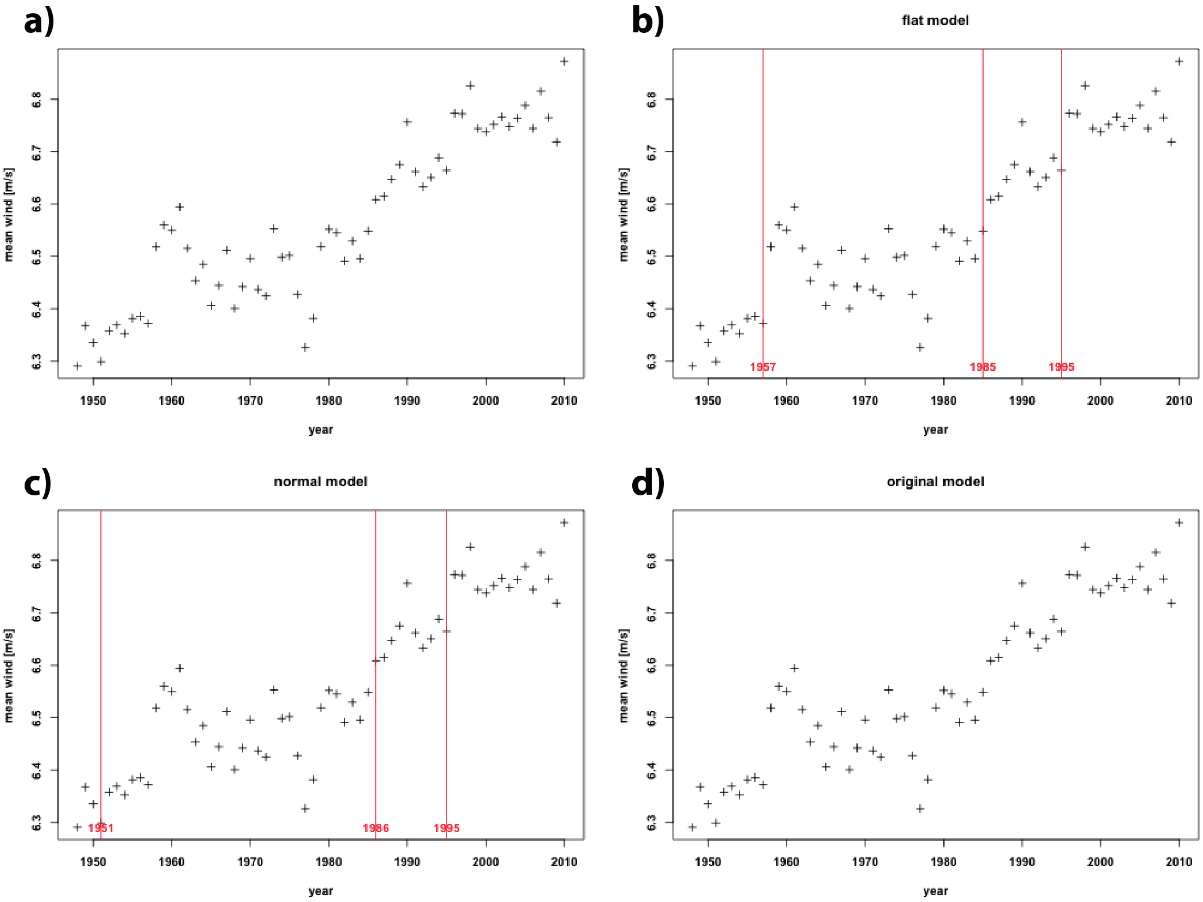


Figure 4.7: Analysis of the results of the block-wise yearly mean wind speed of the global NCEP reanalysis and application of the different modifications of Dose and Menzel. In the upper left, the time series is shown without applying a change point method to it. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right, the change points of the original model are shown.

conducted in these years. It is also visible in the number of observations used to assimilate within the model (Bromwich and Fogt [2004]). Especially the number of used radiosondes shows crucial changes in 1957 and 1979 (Kistler et al. [2001]). Information concerning the consistency after 1996 is not available, since the last published counts of observations for the reanalysis end in 1998 (Kistler et al. [2001]). In this publication, it is not possible to detect an obvious change in the background data for this period.

Next, the influence of the detected changes will be pointed out. Therefore, a differential plot of the raw data is presented in figure 4.6. The difference of the annual mean surface windspeed is plotted for the difference between 1950 and 2010. Values in red indicate years, where the windspeed is stronger in the year 2010. The blue ones show, where the windspeed is higher in 1950. First of all, it becomes obvious, that larger changes occur in the southern polar region. It is also possible to recognise the continents by higher values in the surface windspeed in 1950. The differences of the annual mean windspeed are up to $8\frac{m}{s}$. Summarising, the figure 4.6 shows, that the changes within the variable surface windspeed are of relevant size.

In section 3.2.2, a method was presented, which block-wisely investigates some statistical parameters instead of the whole distribution. The next step is to compare the results of this approach with the ones

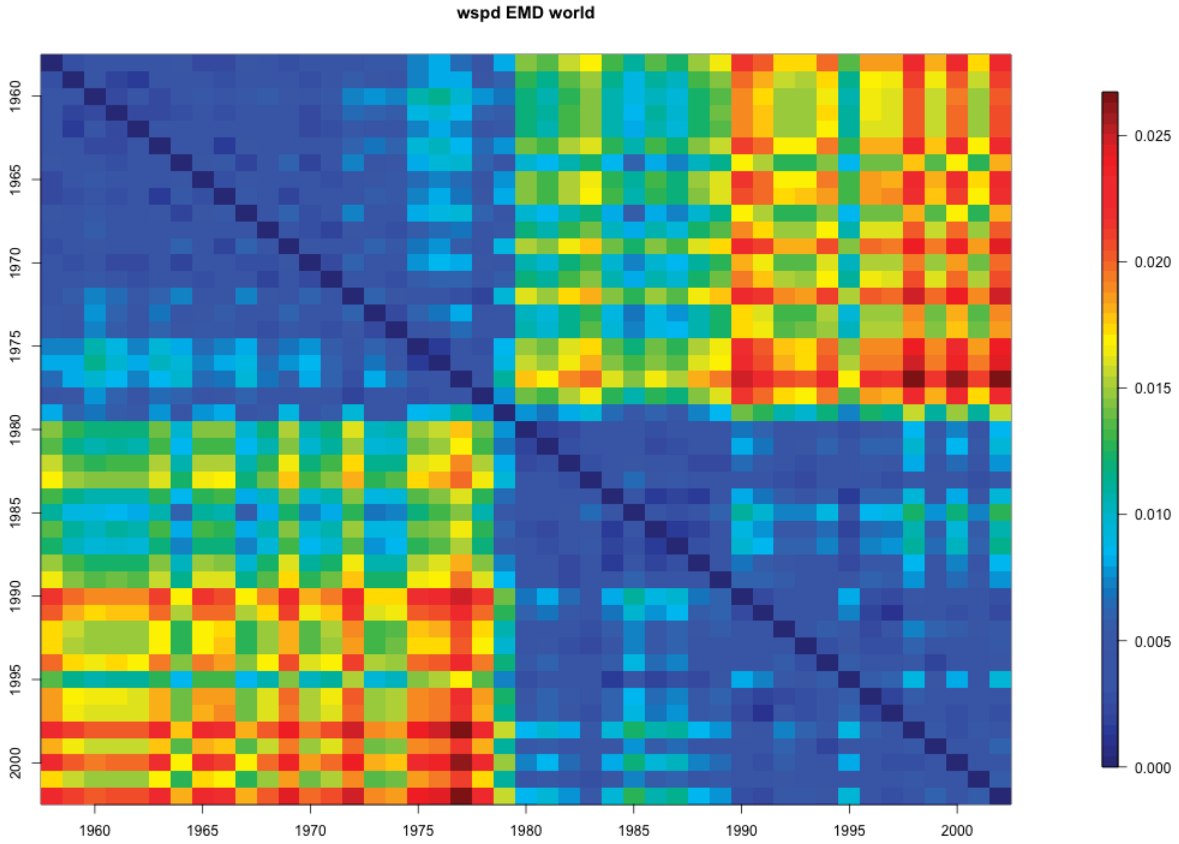


Figure 4.8: Application of the histogram test with the EMD measure to the global ERA40 reanalysis of the surface wind. The x- and y-axis show the years, the values of the matrix are the measured differences by the EMD measure between the histograms of the years.

calculated by the histogram test. Therefore, the block-wise mean of the same dataset is analysed. The mean is taken for every year from all data of the two-dimensional field. This delivers one value for every year, where the results are shown in figure 4.7. Figure 4.7a shows the time series of the mean surface wind of the NCEP reanalysis data in $\frac{m}{s}$. For the whole time series an increase from around $6.3\frac{m}{s}$ to $6.8\frac{m}{s}$ can be determined over the investigated 62 years. In the other three figures, the three main modifications of the Dose-Menzel methods are applied to this time series. The flat model result in figure 4.7b detects three breakpoints in the years 1957, 1985 and 1995. The same results are provided by the histogram test using the EMD measure. Similar results are generated by the normal method, which detects the breaks at 1951, 1986 and 1995. It is shown in figure 4.7c. The EMD equivalent shows similar results, but moved the 1986 breakpoint to 1985. Completely different results are delivered by the original model in figure 4.7d, which does not detect any breakpoint in the mean wind.

For the flat and the normal model the results are the same or at least very similar, while for the original model strong differences exist. At this point, it is not possible to say, which method performs better, if differences in the results exist. All these breaks need further analysis by a data controller, who is familiar with the exact proceedings of the creation of the dataset. Only such a person would be able to say, if the possible physical reasons for these breaks are really represented within the dataset. In the next section,

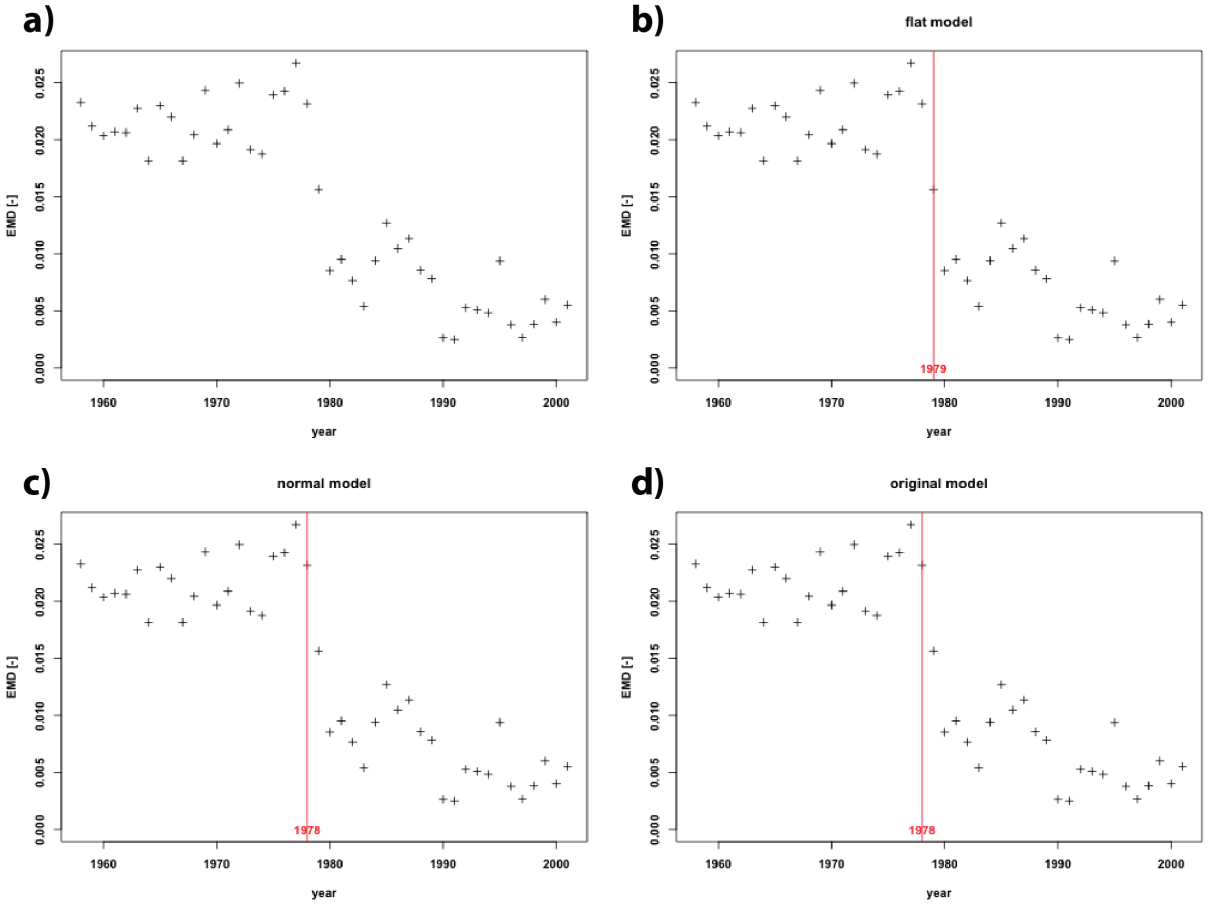


Figure 4.9: Analysis of the results of the histogram test with the EMD measure in comparison to the year 2010 on the global ERA40 reanalysis surface wind speed and application of the different modifications of Dose and Menzel. In the upper left, the raw time series is shown. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right the change points of the original model are shown.

the same methods will be used to investigate a similar dataset.

4.2.2 ERA40 reanalysis

The ERA40 dataset was generated under the guidance of the European Centre for Medium-Range Weather Forecasts (ECMWF). This reanalysis covers the years 1957 to 2002. It is based on the ERA15 reanalysis, which was performed for the years 1979 to 1994 (Uppala et al. [2005]). Later on, it was updated and here the data are used up to the year 2010. Like for the NCEP reanalysis, the assimilated types and number of observations vary in time, so that inconsistencies have to be compensated by the used model.

For the data of ERA40, a similar dataset like in the NCEP reanalysis will be analysed. The data is available as a monthly two-dimensional field for the years 1958 to 2010. Here, also the histogram with the EMD measure is applied. To determine the size of a block, the solution chosen for the NCEP reanalysis is used again. Therefore, all the data for every year is collected from the monthly, two-dimensional fields and the histograms are calculated. The comparisons are shown in the result matrix in figure 4.8.

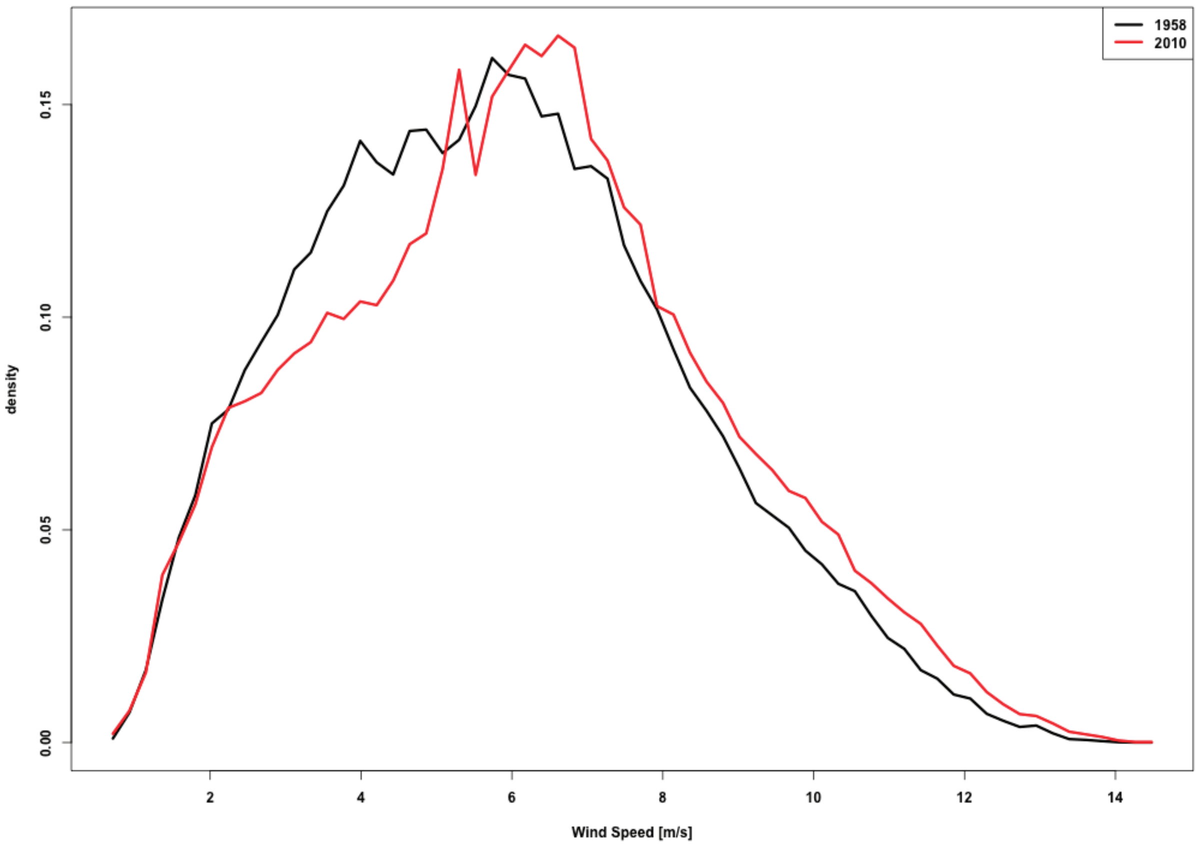


Figure 4.10: Histograms of the ERA40 reanalysis data of the years 1958 (black) and 2010 (red). On the x-axis the wind speed categories in metres per second, on the y-axis the density of each category is shown.

On the x- and y-axis the years corresponding to the blocks are shown. The values of the matrix are the measured distance by the EMD between the dedicated histograms. At first glance, a pattern which indicates a break in the data around the year 1979 can be recognised. Smaller patterns can be identified in the two homogeneous sections, but are not as apparent as the huge break. The combination of these results with the Dose-Menzel method are shown in figure 4.9. Subfigure 4.9a shows the comparisons of all years with the block of 2010. Here, the time series can be divided into two sections, like the visual inspection of the result matrix has shown. The same results, if the three main modifications of Dose-Menzel are recursively used to detect multiple change points. Then, the flat model, shown in figure 4.9b, indicates a break in the year 1979. The two other models (normal model in figure 4.9c, original model in figure 4.9d) set the breakpoint to the year 1978.

The reason for this break can be identified with the introduction of satellite data into the data assimilation of the ERA40 reanalysis (Uppala et al. [2005]). That this change leads to changes within the ERA40 dataset, especially in the southern hemisphere, was also analysed by (Bromwich and Fogt [2004]).

Another question is the influence of the detected break to the dataset. Therefore, the years 1958 and 2010 are compared. The difference measured by the EMD between those two years was 0.023 and, as a consequence, part of the area with relatively high values within the result matrix. The histograms

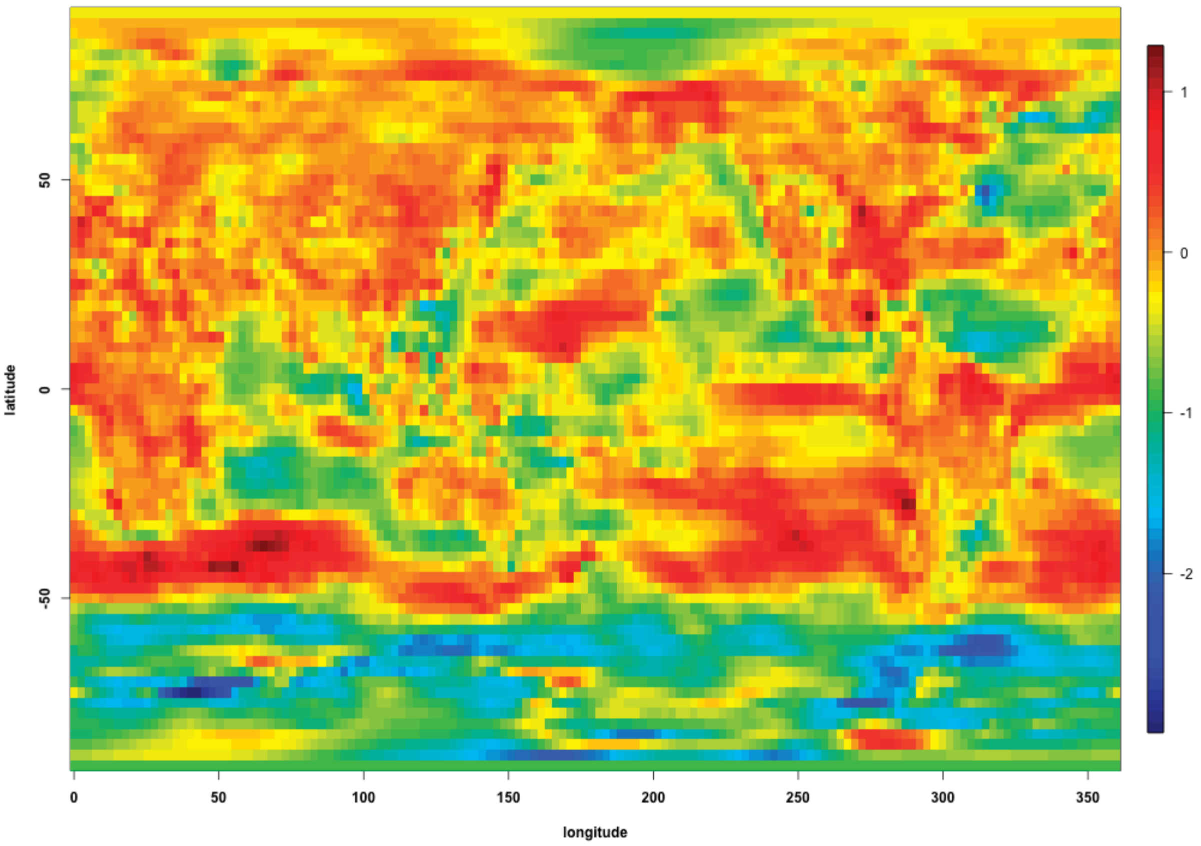


Figure 4.11: Global difference of the yearly mean wind speed between the years 1958 and 2010, measured in metres per second. Positive values in red indicate higher values for windspeed in 2010, negative values in blue indicate higher values in 1958.

compared for this result are shown in figure 4.10. It shows the distribution of the wind speed within the monthly fields of 1958 in black and of 2010 in red. Changes occur for all classes of windspeed higher than $2\frac{m}{s}$. Especially values in the section with a average wind speed of 2 to $5\frac{m}{s}$ are more represented in the data of 1958 than in 2010. The lost density of these windspeed categories is then more or less uniformly distributed to the higher classes above $6\frac{m}{s}$ in the histogram of 2010. Also, the peak has shifted in the dataset. In 1958, it was located at $5.7\frac{m}{s}$, in the year 2010 at $6.6\frac{m}{s}$.

To take a look at the spatial distribution of these changes over the whole globe, figure 4.11 shows the difference of the mean of the surface windspeed of the years 2010 and 1958. Positive values in red represents higher wind speeds in 2010, negative values in blue indicate higher wind speeds in 1958. Apparently, higher values for 2010 can be found in the southern west wind drift zone and on the continents. Higher values for 1958 are mainly located in the southern polar region.

That the change of the wind in the histograms can be visualised without the full application of the histogram test, can be seen in figure 4.12. Here, on the x-axis the years and on the y-axis categories of wind speeds, which are distributed uniformly between the minimum and maximum values of the dataset, are shown. The histograms for the given categories were calculated for the yearly data. The mean of the yearly wind speed is shown in white, the median in black. It is even possible to see the change in the

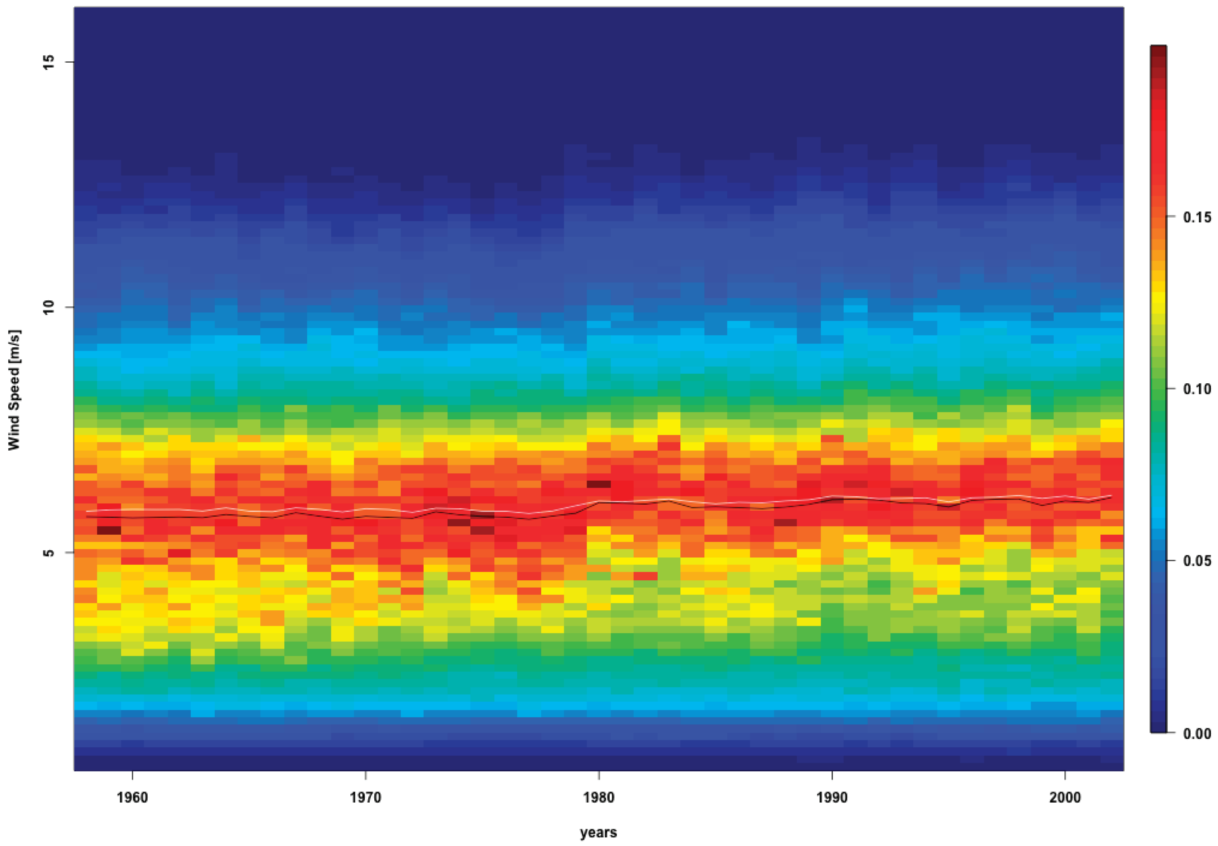


Figure 4.12: Densities of the histograms for each year of the ERA40 reanalysis monthly surface wind speed data. On the x-axis the years, on the y-axis the wind speed categories in metres per second are shown. The white line indicates the mean wind speed for each year, the black line the median.

year 1979 by eye. The years after 1979 do obviously show higher probabilities for higher windspeed. The plot also shows, that the choice of 1958 and 2010 is representative for the different wind regimes within the datasets.

Like for the NCEP reanalysis in the last section, the results will also be compared to the ones resulting from the calculation of a block-wise mean. The latter is calculated by taking the mean of all data of the two-dimensional field of one year and calculating it for all years available. The results are presented in figure 4.13, where subfigure 4.13a shows the time series of the mean wind. Just like the results of the histogram test with EMD applied to the same dataset, shown in figure 4.9, it is divided into two sections. The application of the three main modifications of Dose & Menzel delivers different results than the application to the time series extracted from the result-matrix of the EMD. The results for the flat model are shown in subfigure 4.13b. It sets the breaks to the years 1979 and 1989. The latter break was not detected in the framework of the EMD. For the normal model in 4.13c, the breakpoint was set to 1977, which is one year earlier than indicated by the EMD. Subfigure 4.13d does not show a breakpoint at all for the original model. Reasons can be seen in the construction of the regression of this model, which connects the two parts of the regression by its methodology. A comparison of a constant, linear

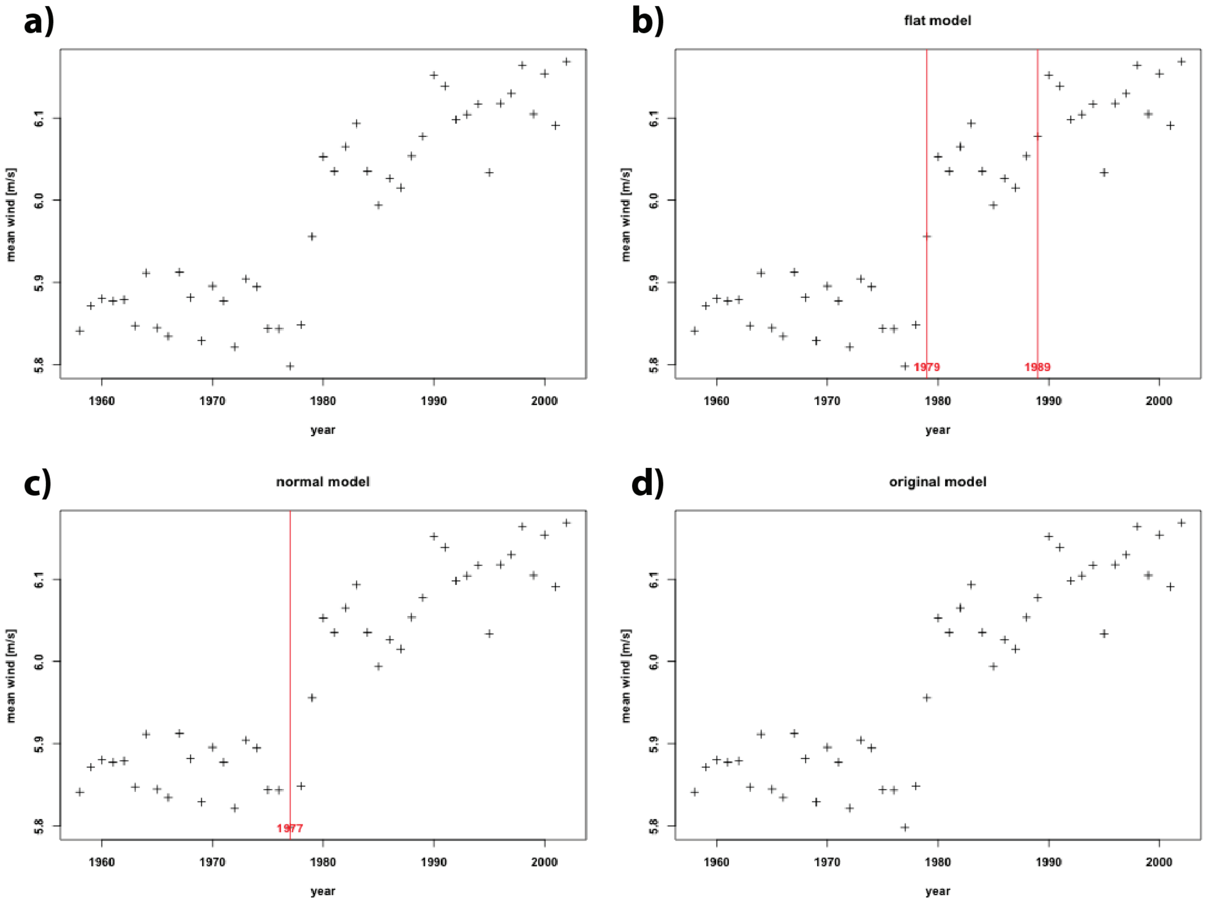


Figure 4.13: Analysis of the results of the block-wise yearly mean wind speed of the global ERA40 reanalysis and application of the different modifications of Dose and Menzel. In the upper left, the raw time series is shown. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right, the change points of the original model are shown.

and change point model with the original regression model shows that a linear regression is favoured. The constant model follows with a big gap. Several orders of magnitudes lower follows in third place the probability of the change point model. Since only the constant and the change point model are compared, the process does not decide to detect a change point.

Again, there cannot be drawn any conclusions about which analysing methods performs best. It can only be seen, that different change points are detected for the mean wind analysis and the EMD analysis and for the different modifications of the change point detection method. Indications for the different change points are given above, but all these have to be checked in detail. Only then, it would be possible to draw conclusions, that proof that the basic data, the ERA40 reanalysis, have strong inconsistencies within the surface wind parameter. This analysis can be done for example within the quality evaluation system, which uses the knowledge and expectations of experts to formalise the conclusions.

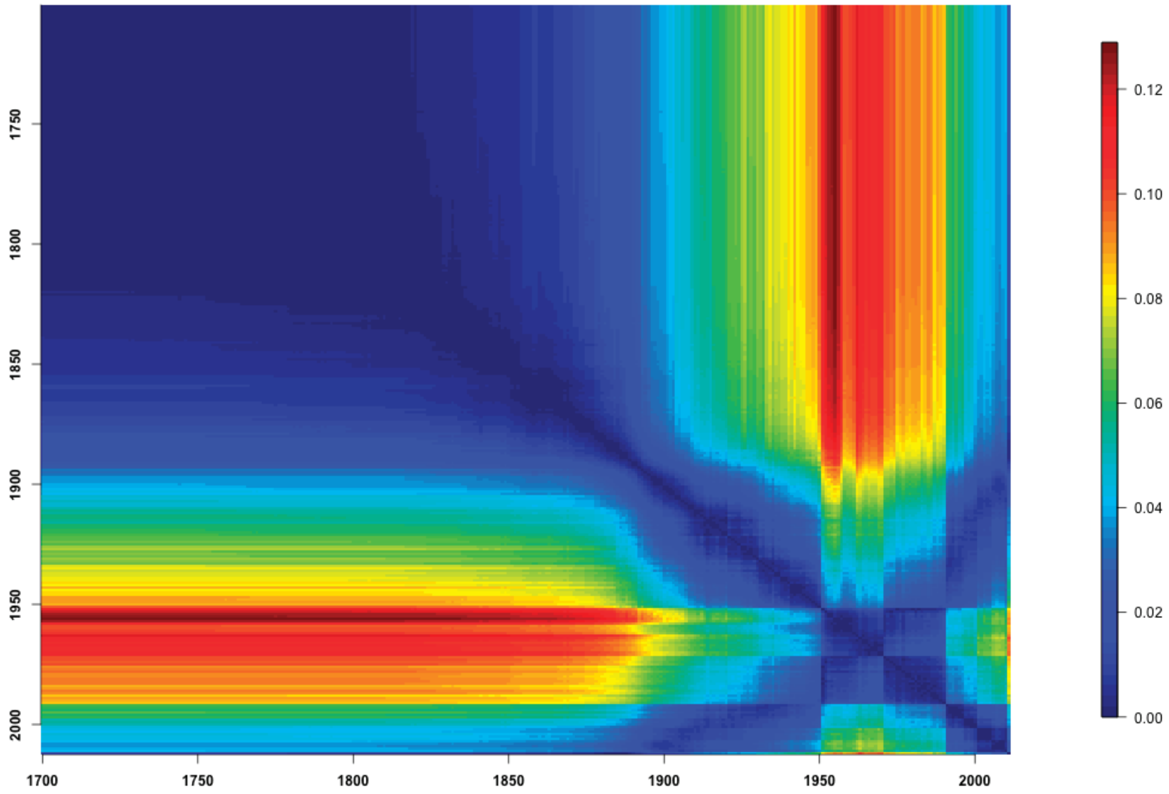


Figure 4.14: Application of the histogram test with the EMD measure to the monthly basis data of the CRU temperature reconstruction. On the x- and y-axis the years are shown, the values indicate the results of the comparison of the histograms.

4.3 Consistency of CRU data

The temperature reconstruction by the Climate Research Unit of the University of East Anglia in the United Kingdom was published in its third edition (HadCRUT3) in 2006. The temperature reconstruction starts in 1850 and is actualised on a regular basis (Brohan et al. [2006]). In July 2011, a larger subset of the basic data for this reconstruction was published by the UK Met Office. It consists of the most of the originally used 4349 stations (Met Office [2011], Brohan et al. [2006]). The temporal resolution for both the basic data and the reconstruction is monthly.

In this analysis, the basic data will be evaluated first. Therefore, the histogram test with the EMD measure is applied to the dataset. The blocks are built by the collection of all temperature data of the available stations for one year. Then, the histograms and the comparison are calculated. The result matrix is shown in figure 4.14.

The plot shows the matrix resulting from the data available between 1700 and 2011. On the x- and y-axis the years can be found. The values within the matrix are the results of the comparisons measured by the EMD measure. Obviously, the values increase slowly from 1890 to 1950. Afterwards, a pattern, which delivers a cut in the results, is recognisable. This happens again with a small inconsistency in 1970s and

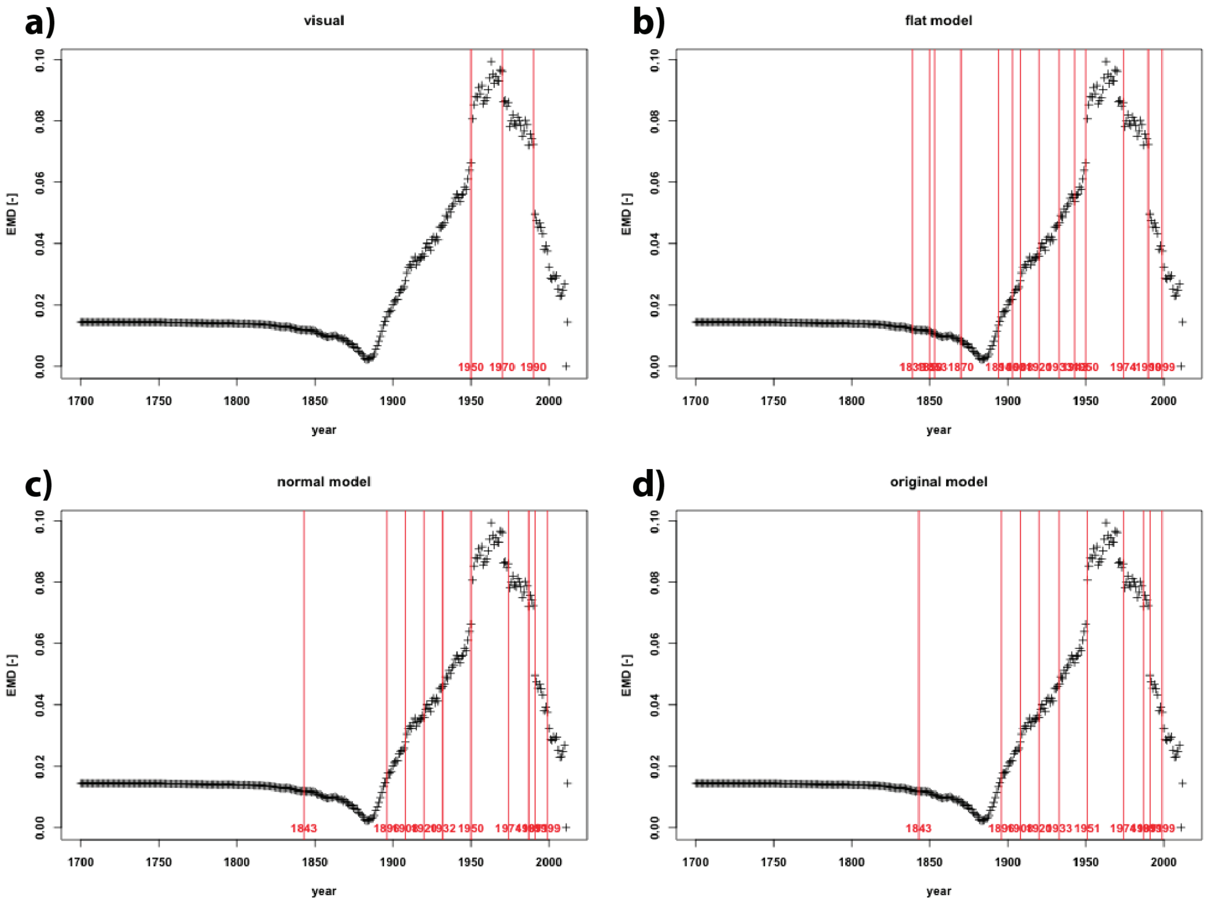


Figure 4.15: Analysis of the results of the histogram test with the EMD measure in comparison to the year 2010 on the basis data of the CRU temperature reconstruction and application of the different modifications of Dose and Menzel. In the upper left, the time series is shown without applying a change point method to it. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right, the change points of the original model are shown.

again with a stronger one in the 1990s. Afterwards, the data seem to be consistent until 2011. As the last year was not fully represented in the underlying data, it behaves different from the years before. For the combination of the method of Dose-Menzel and the histogram test, the comparisons with the data of the year 2010 are chosen. For 1700 until 2009 they are shown in subfigure 4.15a. Included are the obvious inconsistencies, detected by visual inspection, which were mentioned before. This leads to the breakpoints in the years 1950, 1970 and 1990. On this data, the main modifications of Dose-Menzel are applied recursively. In the other subfigures of figure 4.15 the breakpoints are shown for the flat model (figure 4.15b), the normal model (4.15c) and the original model (4.15d). In all three casesm the number of detected breakpoints is higher than the one identified by the visual inspection. The flat model indicates 14 breakpoints (1839, 1850, 1853, 1870, 1894, 1903, 1908, 1920, 1933, 1943, 1950, 1974, 1990 and 1999). Most of them are included due to the fact, that the increase and decrease of the values have to be regressed by a horizontal line. Therefore, the number should be lower for the other two models. Indeed, both deliver only ten breakpoints. They are placed similarly in these models. The normal model puts them to 1843, 1896, 1908, 1920, 1932, 1950, 1974, 1987, 1991 and 1999. The original model skips the

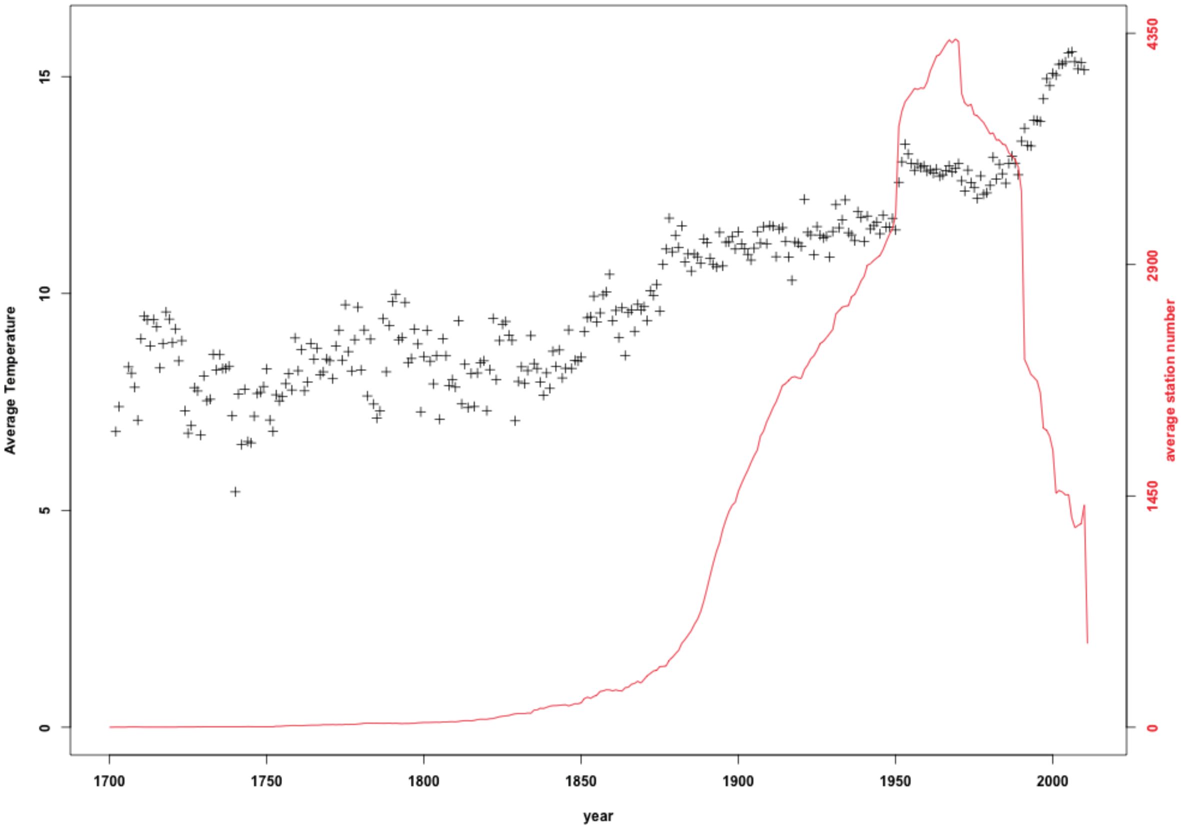


Figure 4.16: Average temperature and average number of station per year of the basis data of the CRU temperature reconstruction between 1700 and 2010. On the x-axis the years are shown. The left y-axis indicate the temperature in degrees of Celsius, while the right one shows the number of stations.

1932 and 1950 break events and sets both to one year later. Both cases find breakpoints at the increasing and decreasing parts of the values, where little breaks are fitted by the inclusion of new sections. The three breakpoints detected by the visual inspection are nearly found by all models. The break in 1970 is fitted four years later by all models, 1950 only one year later by the original model. The normal and original model also replaced the break from 1990 to two breaks in 1987 and 1991. Summarising, it can be said that the methods react sensitive, but deliver acceptable results. The most important conclusion is, that they are able to indicate the inconsistency of the data.

Reasons for the inconsistencies can be found by taking a look at the number of used stations in each year. This is done in figure 4.16. In this plot, the yearly average temperature of all measuring stations is plotted as crosses. On the right y-axis, the yearly mean number of stations is shown as a red line. The breakpoints found by the models and the ones seen here in the mean temperature, are accompanied by changes of the number of stations.

An arising question is, how these inconsistencies of the input data affect the results of the temperature reconstruction. Additionally, it shows how well the applied methods, described by Brohan et al. [2006], can deal with the problematics of inconsistent input data. Therefore, the temperature reconstruction,

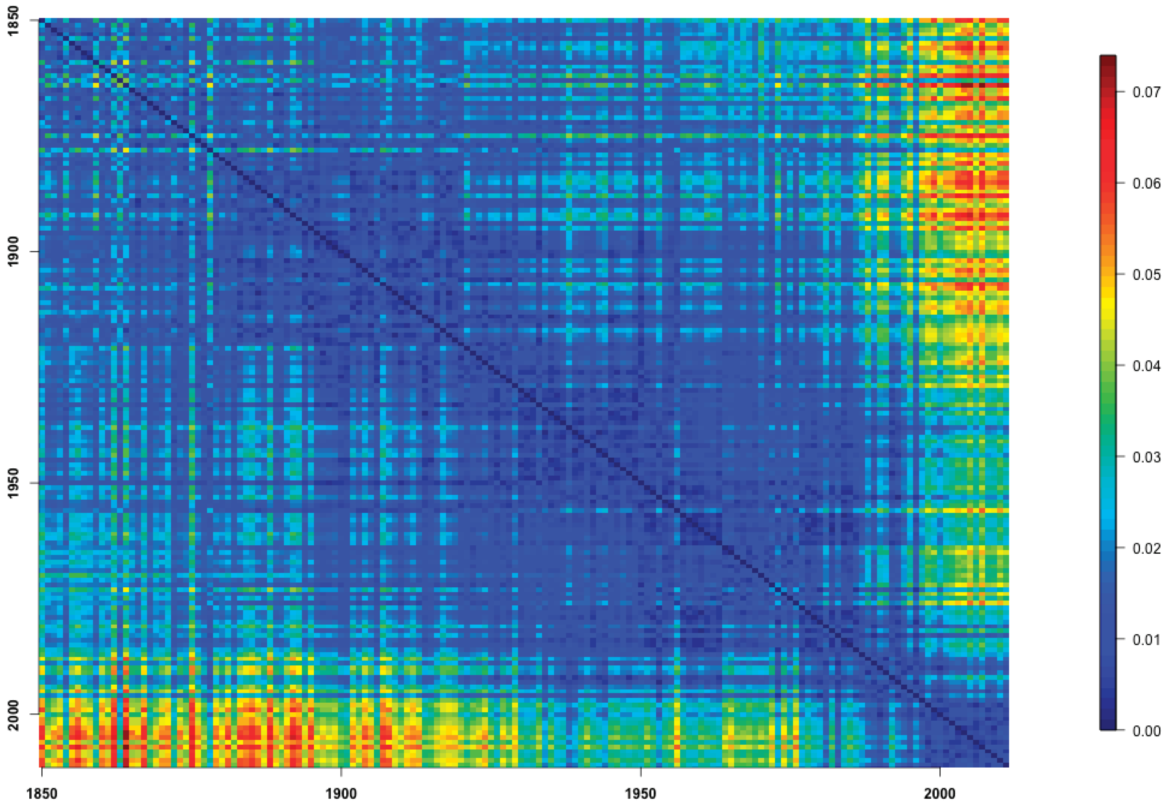


Figure 4.17: Application of the histogram test with the EMD measure on the global CRU temperature reconstruction for the years 1850 to 2010. The x- and y-axis show the years, the values of the matrix are the measured differences by the EMD measure between the histograms of the years.

ranging from 1850 to 2011, is also analysed by the histogram test with the EMD measure. The blocks for the comparison have to take the two-dimensional surface temperature fields of one year into account. Therefore, all values at each grid point for every month of one year are included into one block. From these blocks the histograms are calculated and compared. The result matrix is shown in figure 4.17, where the x- and y-axis represent again the years of the blocks. It shows, that the vast majority of the plot consists of relatively low values. It only changes for the comparisons with the years after 1990, where relatively high values can be found.

Just like for the raw dataset, the year 2010 is taken as a reference for the combination of histogram test and the method of Dose and Menzel. The extracted time series is shown in figure 4.18a. It shows that for a long time span from 1850 to the 1920s the distance is relatively constant. This is followed by a section with lower values up to the 1970s, which is marked by some relatively high values. Afterwards, the values decrease steadily. The application of the flat model (figure 4.18b), the normal model (4.18c) and the original model (4.18d) to the data, delivers three to four breakpoints. The flat model sets the first change point to 1924, the second to 1976 and subdivides the decreasing section by including a break to 1996. The normal model includes one more change point and places the others slightly different. The

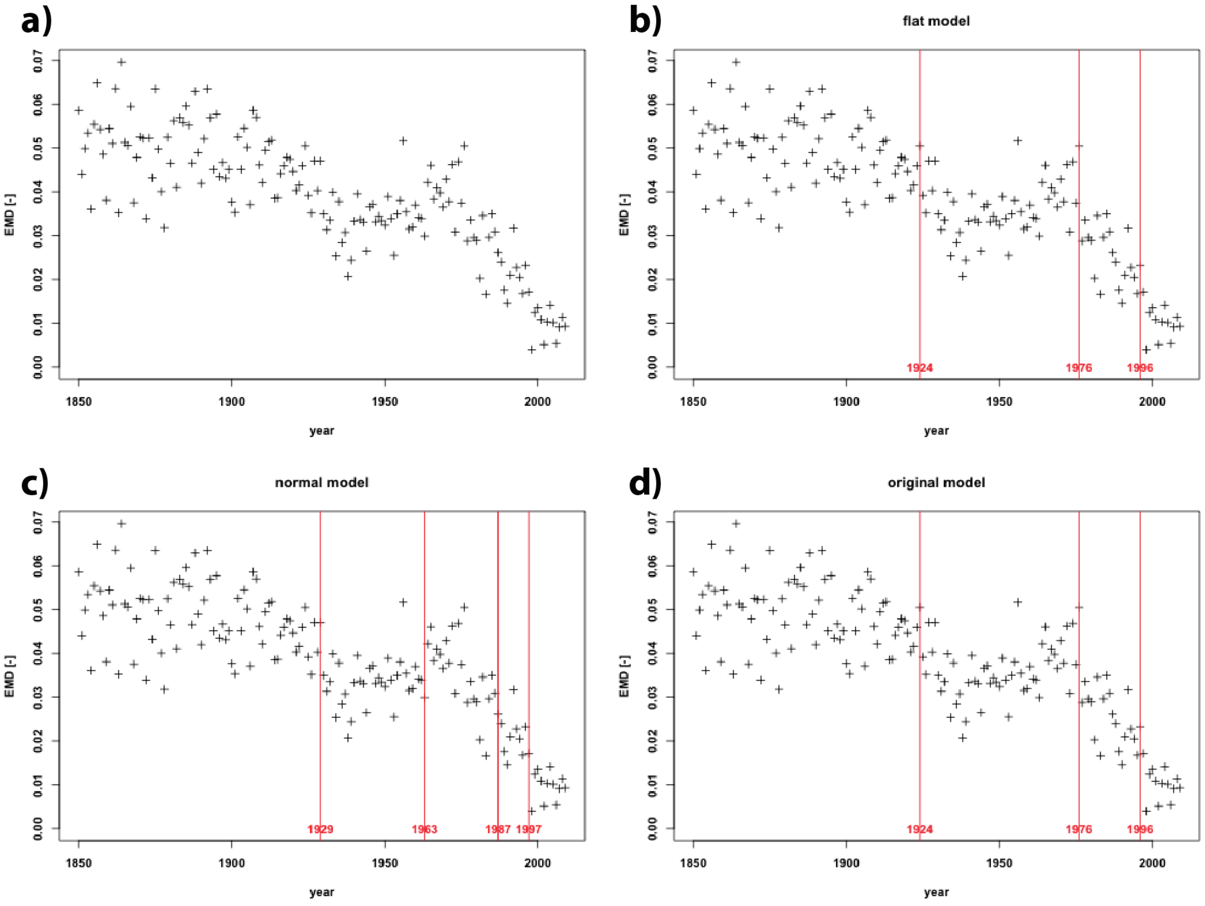


Figure 4.18: Analysis of the results of the histogram test with the EMD measure in comparison to the year 2010 on the global CRU temperature reconstruction and application of the different modifications of Dose and Menzel. In the upper left, the time series is shown without applying a change point method to it. In the upper right, the detected change points by the flat model, in the lower left those of the normal model and in the lower right, the change points of the original model are shown.

first is positioned at 1929, which is five years later than the flat model. The following section ends in 1963, before the decreasing section has two more change points in 1987 and 1997. The original model delivers again the same breakpoints like the flat model.

All these change points seem to be reasonable, even though a visual inspection would not have detected all of them. Especially the pattern after the change point at 1996/1997 would be expected under an assumed global warming within the dataset.

Table 4.3: Measured variables at the climate station in Hohenheim (Wulfmeyer and Henning-Müller [2007]).

Abbreviation	Meteorological Parameter	Unit	Time [s]
TimeFromStart	Measurement time	s	30
T200	Air Temperature 2 m above ground	°	30
RH200	Relative Humidity 2 m above ground	%	30
T005	Air Temperature 5 cm above soil	°C	30
WD1000	Wind from Direction 10 m above ground	Degree	30
WS1000	Wind Speed 10 m above ground	m/s	30
GR200	Global Radiation 2 m above ground	W/m2	30
RR200	Reflected short Radiation 2 m above ground	W/m2	30
NR200	Longwave Radiation Budget 2 m above ground	W/m2	30
RAIN	Rain-Collector 1 m above ground	mm	30
ST002	Soil Temperature at 2 cm under bare soil	°C	30
ST005	Soil Temperature at 5 cm under bare soil	°C	30
ST010	Soil Temperature at 10 cm under bare soil	°C	30
ST020	Soil Temperature at 20 cm under bare soil	°C	30
ST050	Soil Temperature at 50 cm under bare soil	°C	30

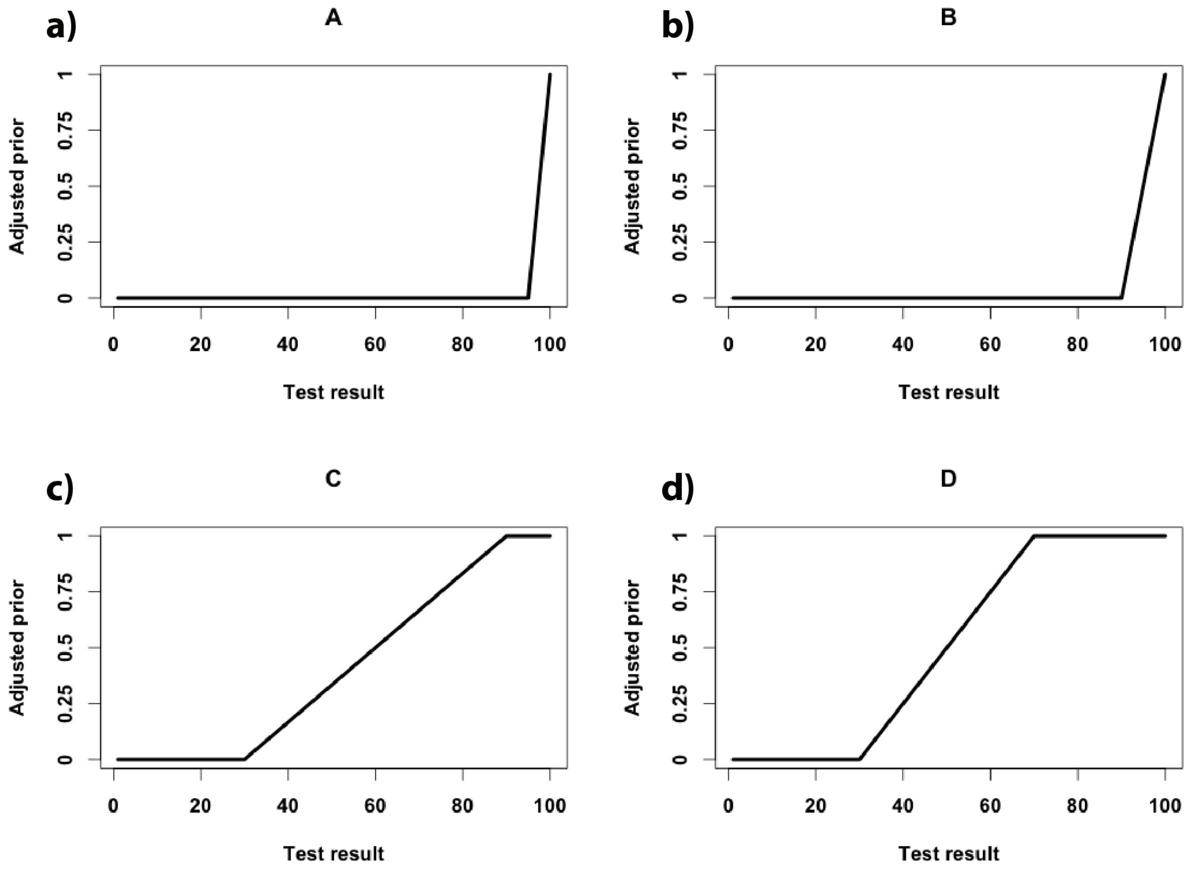


Figure 4.19: Priors used in the quality evaluation of the climate station in Hohenheim. On the x-axis the test results of the quality tests and on the y-axis the dedicated adjusted prior are shown. The priors range from very strict (A) to relatively loose (D).

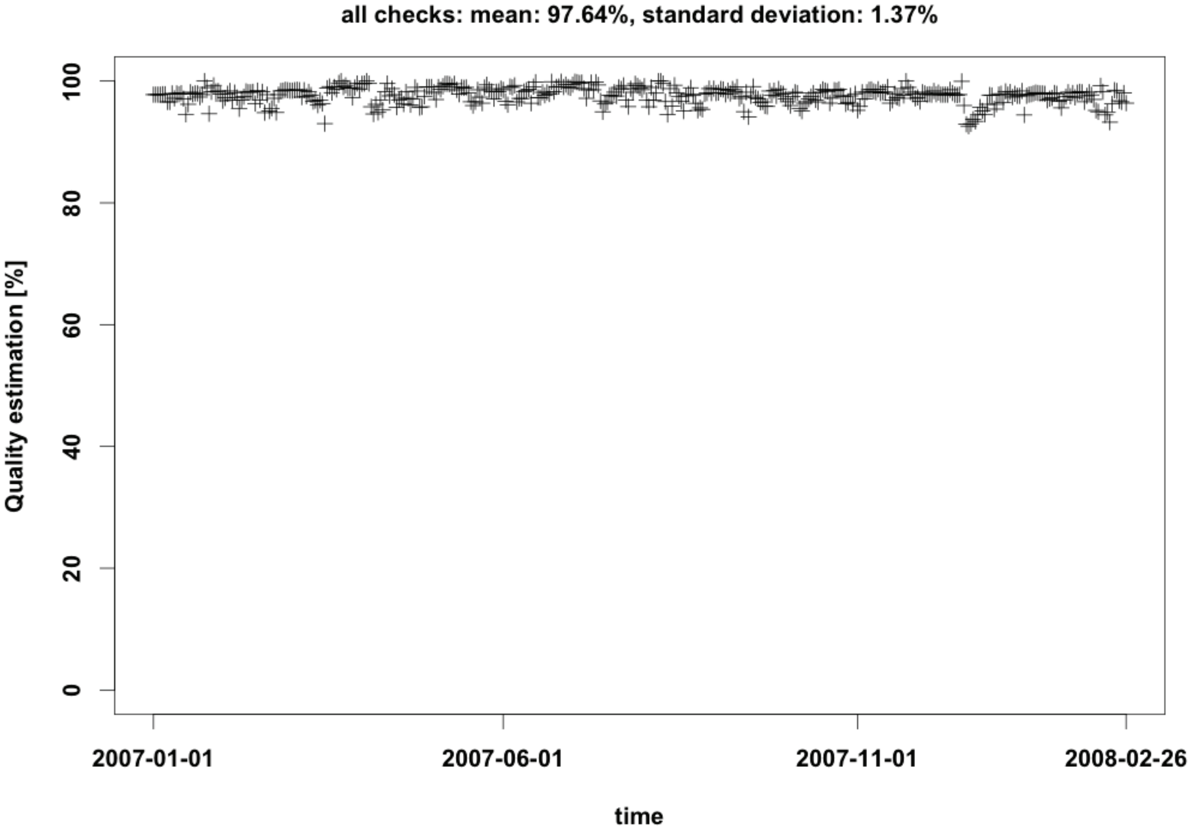


Figure 4.20: Overview of the quality estimation for every day and for all meteorological variables of the climate station Hohenheim. The x-axis indicates the days, the y-axis the quality estimation in percent.

4.4 Quality estimation of Hohenheim climate station

In a last application, the quality estimation process, described in section 2.6.7, will be used to evaluate measurements of a climate station. The station is located at the University of Hohenheim and was part of the Global Observation Period (GOP) within the DFG priority program "Quantitative Precipitation Forecast" (Hense and Wulfmeyer [2008], Crewell et al. [2008]). Available is the time span from 1st January 2007 to 26th February 2008, and as a consequence 421 days of data. Measured by the station are 15 variables, which are shown in table 4.3.

4.4.1 The setup of the quality evaluation system

Before the testing procedures start, which evaluate the quality, several parts have to be defined. The first are the tests, which should be performed on each variable. In this application, the LIM, ROC and NOC tests by Meek and Hatfield [1994], which were described in section 3.1, are used with different parameters for each variable. Since the tests do not automatically deliver probabilities, like it is required by the used quality evaluation system, the determination of the probabilities has to be explained. The result of each of the tests is a flag vector. This vector contains 0s for values, which are not classified as suspicious by

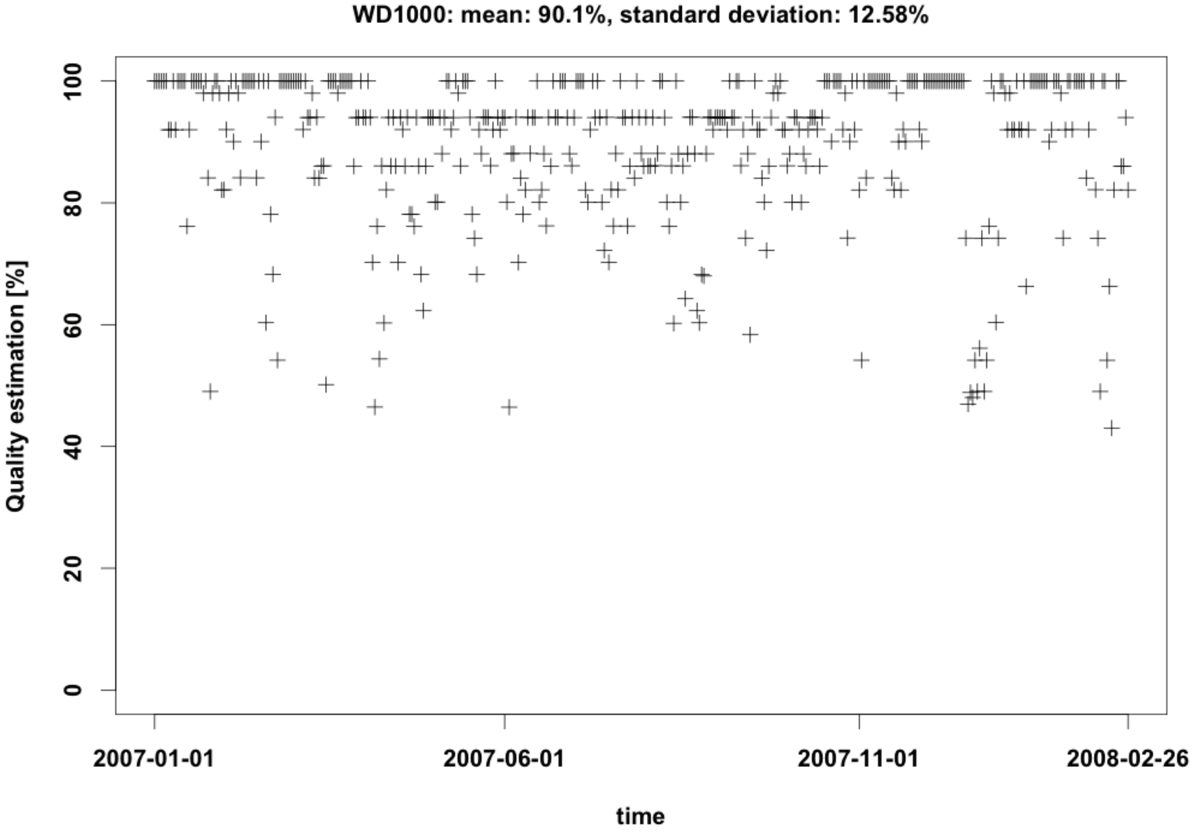


Figure 4.21: Quality estimation for every day for the wind direction measured by the climate station Hohenheim. The x-axis indicates the days, the y-axis the quality estimation in percent.

the test, and 1s for the others. Therefore, it is possible to calculate the percentage of unsuspicious values for each test. This can be seen as the quality estimate by this test under the given conditions.

Other quantities to be defined are the general priors and the weightings. Basically, four different priors are used. They are shown in figure 4.19 and range from a strict requirement (A) to pass the test, to relatively loose requirements (D). Prior A (subfigure 4.19a) is characterised by a fast decrease from 1 to 0, when the result of the test fails to gain 100%. The second prior B (subfigure 4.19b) is similar, but has a slower decrease. For the third prior C (subfigure 4.19c), the first ten percents probability for failing the quality test still give an adjusted prior of 1. Afterwards, the prior decreases linearly until the probability of succeeding the quality check falls below 30%. Prior D (subfigure 4.19d) has a similar structure, but allows an adjusted prior of 1 for the first 30% probability of failing the test.

The weighting is chosen as the follows: For each of the variables, all used parameter sets are assigned a weight from 1, for tests with low importance, to 3, for highly important tests. To obtain the effective weight for each test result, the specific weight for the test is divided by the sum of all weights that belong to all tests of the corresponding variable. To get estimations for the whole dataset with all variables, the quality result for each variable is weighted equally.

The used tests, parameters, priors and weights are shown in the appendix in the tables B.1 and B.2. In the following section, the results will be shown and some suspicious events will be further investigated.

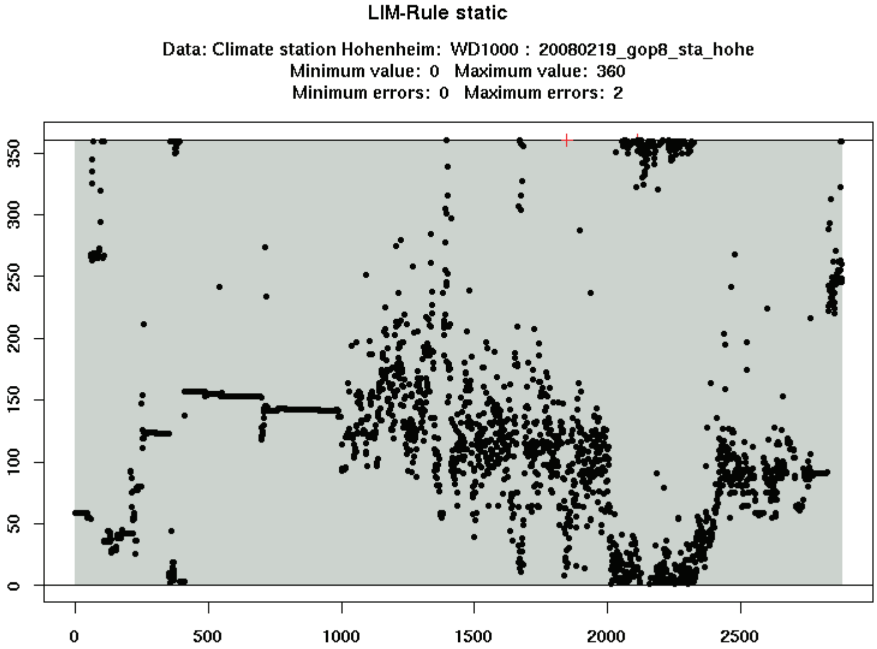


Figure 4.22: Application of the LIM static test on the wind direction data of the climate station Hohenheim at the 19th February 2008. The x-axis shows the number of values in the dataset, the y-axis the wind direction measured in degrees. The maximal limit of the test is set to 360, the minimum limit to 0. The region in between has a grey shading and the values located in this range are marked as black dots. Values outside have red crosses.

4.4.2 Results

As a first result, the quality estimate for the whole dataset with all variables is illustrated in figure 4.20. The x-axis shows the time for every daily dataset and the y-axis the mean quality estimate of all variables for each day in percent. Generally, the values are very high and no certain pattern is recognisable. The mean of the quality estimates is 97.6% and the standard deviation is 1.4%. Depending on the chosen prior information and tests, it can be concluded, that the dataset has a high quality. Nevertheless, the values of the quality estimation suggest, that it is not perfect for any day. To take a look on the reasons for the missing percentages, two examples are shown in the following. The quality estimate for the wind direction in 10 metres height is illustrated for all days in figure 4.21. Here exist several days with a low quality. The mean of the quality estimate is with a value of 90.1% lower than the estimate for all variables together. The standard deviation of 12.6% indicates a high variability of the estimated quality. To look for the reasons of these low values, the day with the lowest value, reached with 43,0% at the 19th of February 2008, is further investigated. The plot of the LIM test static applied to the wind direction for this day is shown in figure 4.22. The used parameters are the minimum limit of 0 and a maximum limit of 360 degrees. On the x-axis, the number of values of this dataset is shown. Since the measurement takes place every 30 seconds, it shows 2880 data values.[6] On the y-axis, the values of the measured wind

[6] The original datasets in the file consists of 2881 values. Since the last value was in several order of magnitude higher than the other values, it is interpreted as an error value and not considered in the evaluation here.

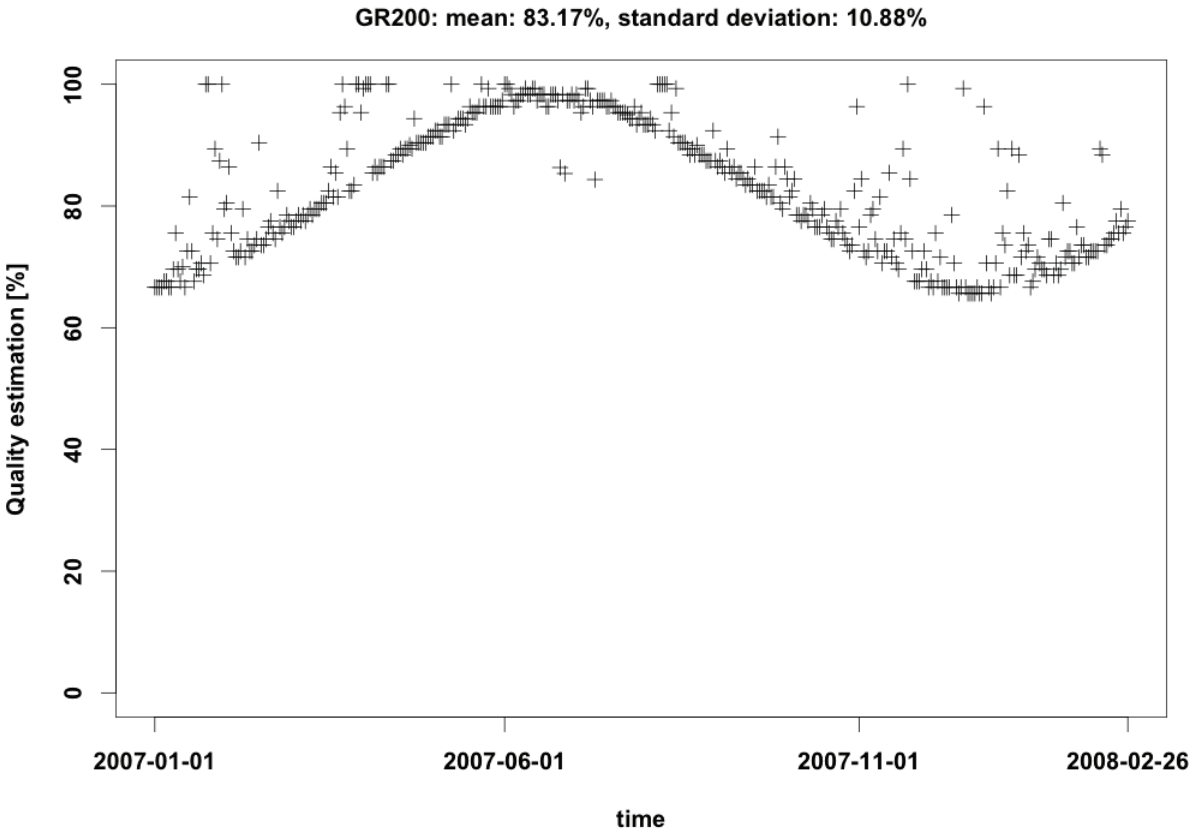


Figure 4.23: Quality estimation for every day for the global radiation in 2 meter height, measured by the climate station Hohenheim. The x-axis indicates the days, the y-axis the quality estimation in percent.

directions are shown. From the test results that two values are above the given limit of 360 degrees.
An expectation concerning the wind direction is, that it is limited by 0 and 360 degrees. As a consequence, the test was given a high weighting and a strict prior. Additionally, four NOC tests are performed on the dataset, at which they detect several hundred events. An example is the NOC test, which accepts 20 consecutive elements with the same value as a maximal number of repetitions and finds 278 repetition errors. Since this has the same weighting and prior as the LIM static test, it extremely lowers the quality estimate of this variable. This finding makes it simple to find and analyse the potential errors in the dataset and allows the controller of the datasets to comment on them.
Another example is given in figure 4.23. It shows the quality estimation for the same time span for the ground radiation in two metres height. At first glance, a pattern with similarities to a wave becomes apparent. Since the frequency is around a year, the suggested conclusion is that the used test detects an annual cycle within the dataset. An example for the 1st January 2007 underlines this. The result of the LIM static test with the limits 0 and 1367 $\frac{W}{m^2}$ is shown in figure 4.24.[7] It shows, that several values

[7] These limits are chosen, because they allow with the upper limit the possibility to detect gross errors. The lower limit is chosen under the assumtion, that only positive values are possible. Since the range of values depends on the definition of the parameter 'ground radiation' and no further metadata is available, negative values does not mean necessary an error within the dataset.

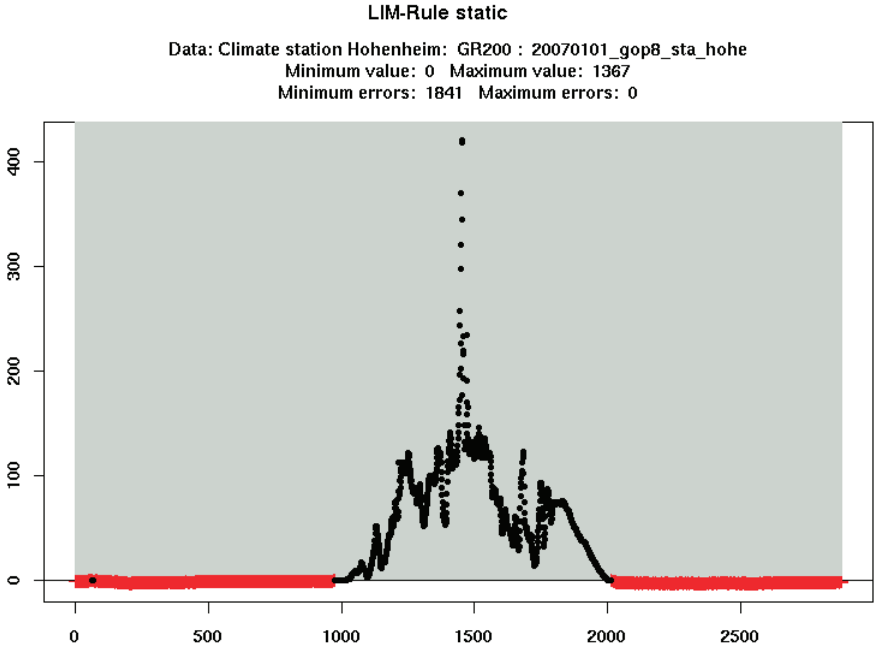


Figure 4.24: Application of the LIM static test on the two metres ground radiation data of the climate station Hohenheim at the 1st January 2007. The x-axis shows the number of value in the dataset, the y-axis the global radiation measured in watts per square metre . The maximal limit of the test is set to 1367, the minimum limit to 0. The region in between has a grey shading and the values located in this range are marked as black dots. Values outside have red crosses.

are below the lower limit and are therefore flagged. Since these negative values only happen at night, their occurrence obviously depends on the length of day. This explains the annular cycle in the quality estimation.

The previous examples have shown the basic functionality of the quality evaluation system and its high dependence on the used tests, parameters and priors of the results. As a consequence, all of them have to be set by people who are real experts both of the dataset and their generation. Further discussion on that topic follows in the next chapter 5.

5 Discussion

In this chapter, the proposed procedures and results will be discussed. Open questions will be addressed, more general ideas explained and some remarks on possible future developments in the field of scientific data publication will be made. The section is divided into three parts. The first part in section 5.1 revisits the proposed general tests. The main focus will be set on the newly developed histogram test, introduced in section 3.3, and the change point detection method by Dose and Menzel [2004], described in section 3.4. The latter was used here for quality assurance purposes for the first time. For both methods, the sensitivity tests are briefly discussed and some remarks on their performance are given.

The second section 5.2 discusses the quality evaluation procedure, which was proposed in section 2.6.7 and in section 4.4 applied to data from a climate station. Here, the focus will be set to the interpretation of the results for such a procedure and which risks it may bear and which chances it may offer. The last section 5.3 concentrates again on the data publication in general, which was already extensively introduced in chapter 2. The proposed working scheme and its consequences for these types of publications will be further investigated. In addition, the integration of the proposed procedures into existing infrastructures will be briefly discussed. The last topic in this section will be a possible effective peer review process for primary data. It will be illustrated, how the developed procedures like the general tests and the quality evaluation could contribute to the unsolved problem of peer review.

5.1 General Tests

The necessity to develop general tests was explained in section 2.6.6 and as a consequence four types of general tests were proposed in chapter 3. In this chapter, all four will be discussed, although the focus will be set on the histogram test and the change point detection method by Dose and Menzel.

The at first proposed tests by Meek and Hatfield [1994] in section 3.1 represent the prototype of general tests, because of their flexibility and simplicity. Nowadays, they are used in a wide range of applications. With the enhancements by the proposed modifications, they can be used in even more applications. One of the applications was shown in section 4.4, where it was used as a general tool for quality evaluation. The method used there to calculate a probability delivers acceptable results, but is far from ideal. Therefore, the priors have to account for problems, like for example even a very low number of values flagged by a test might be an indication of a dataset with a low quality. More discussion on the quality evaluation technique itself follows in section 5.2. Also, the dependence on the parameters requires a very thoroughly performed documentation of the methods to ensure that they can be replicated.

A second class of tests was shown in section 3.2. It presents tests that calculate different statistical parameters by different means. The used statistical parameters are moments and percentiles that allow to gain indications on the whole distribution. Different types of selecting the subsets under consideration have different abilities of detecting inconsistencies. As a consequence, it depends on the application, which type should be chosen. Nevertheless, the information reduction is strong so that several parameters have to be considered to obtain a good indication of the problematics within the datasets under consideration.

This stands in contrast with the third test, the histogram test, which checks the whole distribution at once. The histogram test will be discussed in detail in the next section 5.1.1. Here, the focus will be set

on the discussion of the different distance measures and their abilities to indicate inconsistencies.
The fourth test, discussed in section 5.1.2, describes the abilities of the change point detection method introduced by Dose and Menzel. Besides the general discussion on the sensitivity tests and the intercomparison tests, it will be the combination with other methods, that will be discussed here. This will be done in order to lay the foundation for the integration of further methods into the quality evaluation procedure, which will be discussed afterwards in section 5.2. Furthermore, the method shows abilities for applications within homogenisation procedures, which will also be discussed there.

5.1.1 Histogram test

The histogram test was introduced in section 3.3 and some applications were shown in the sections 4.1 to 4.3. The basic idea to compare blocks of a dataset with each other, is similar to the test, which compares statistical parameters block-wise and was described in section 3.2.2. The main advantage of the histogram test is that it compares at once the whole distribution of a block with the other blocks. Therefore, it uses different distance measures for one-dimensional histograms. These different measures deliver different results, which allow a flexibility in the application of the test, since their sensitivity to different kinds of inconsistencies varies.
The result of the test is a matrix in which every value describes the distance of two histograms, expressed by a distance measure. The result is not as simple to handle as the ones of other quality checks. Even if the user of the test is used to similar matrices, like covariance matrices, he/she will have to learn, how to interpret them. Nevertheless, it is quite intuitive for users to look just for patterns in a matrix when they want to know, if there might be problems within the dataset.
This section will discuss in a first step, in section 5.1.1.1, the sensitivity tests performed in section 3.3.3. The results of the different distance measures will be set into a context, which explains their behaviour. The influence of different measures and their consequences will be discussed in section 5.1.1.2. In a last step, the section 5.1.1.3 looks at the reduction of information performed by the histogram test. Additionally, the fields of applications, especially the use for multidimensional time series analysis, are discussed. The relativity of the results, which may lead to over-interpretation of them, will be briefly discussed at the end.

5.1.1.1 Sensitivity of the method

In section 3.3.3 three different sensitivity tests were shown. One of these tests is used as a calibration for the methods using the Kullback-Leibler and the Jenson-Shannon Divergence. The other two ones are comparison tests of the five presented distance measures used for this procedure.
The first point to be discussed here is the evaluation method. The distinguishing quantity x_{sd} for all three tests was defined in equation 3.18. The basic assumption is that it is known, where the inconsistencies in the dataset starts (see also section 3.3.3.1). From this, it is possible to identify the regions of the matrix, where which types of characteristics are compared in the histogram comparisons. Generally, the method works and the delivered results are valid. It proves to be able to deliver obviously acceptable results, about whether regions of a matrix are distinguishable or not. The alternative to decide only by visual inspection only, is not usable, since it is too subjective and not scalable for tests with a high number of matrix evaluations. The information given by the quantity x_{sd} is the difference between the means of two regions measured in their respective standard deviations. This information, as a general information criterion in the practical work, is only of minor relevance. Nevertheless, the defined limit of $x_{sd} = 1$ to distinguish between two regions is a valid approximation. It should also allow further pattern recognition, what was shown for example with the combination of the histogram test and the change point detection

method. Further discussion on that topic will be given in section 5.1.2.4.

The first sensitivity test was the calibration of the prior a_p, given by equation 3.12 for the KLD and JSD measures. The dependence on the number of observations s_b used in one histogram, is a reasonable assumption of the prior. It prevents the prior from getting too much weight within the histogram to influence the results. The value a_p determines the influence of the empty bins on the results. Since the influence should not be too high, the value is fixed for all further calculations using the two measures. The value of $a_p = 100$ was used, because the choice of a higher value would not have a strong influence on the results. Therefore it is plausible that with choosing a different a_p, the results for the KLD or JSD would not change to the better in the following sensitivity tests. As it is shown in figure 3.2, the resulting x_{sd} also depends highly on the number of bins of the histogram n_b. This should be kept in mind, when the method with the KLD and JSD measures is used in new fields of application.

The method used to calibrate a_p is to investigate, if the method is able to detect a rounding to the first digit for a standard normal distributed vector. The calibration method is not necessarily ideal, but a valid one. Another possibility could be the calibration with a detection of a given shift in mean or variance.

In a second sensitivity test, the detection of a shift in the mean within a vector was evaluated, what was explained in section 3.3.3.3. Apart from the result, that proved the EMD to be the most sensible measure to small shifts in the mean, the general behaviour of the methods is of importance. It shows that the general type of the used distance measure explains the performance of the method. This can be seen from the division of the five methods into three different strains. The first is given by the RMS and MS methods, which both have the same basis. The same holds for KLD and JSD in the second strain. All four measures are defined by a bin-wise calculation of the difference of the two histograms. The third strain, given by the results of the EMD measure, shows a completely different behaviour. This becomes especially obvious for a small shift. An explanation can be given by the different basic mechanism, which is used by the EMD as a solution to an optimisation problem. It can be interpreted as a transportation of probability of one pdf, estimated by the histogram, to the other pdf. In case of a shift in mean, the difference between both histograms has to be transported from one end of a histogram to the end of the other side of the other histogram. As a consequence, this adds up a lot of value to the result of the EMD, which is therefore more sensitive than a bin-wise comparison. Further discussion follows in the next section.

A similar result is shown in section 3.3.3.4. In the third sensitivity test, a shift in variance is evaluated. Here, the EMD also proves to be more sensitive than the bin-wise methods. A finding, which is not shown in the plots, is the stability of the results of the EMD for an increasing number of bins n_b. The other measures reach the detection limit of $x_{sd} = 1$ later, if the number of bins n_b is increased.

With the following example it can be shown that the difference of reaching the detection limit is not an artefact of the evaluation method. Therefore, a vector with a shift in mean of 0.4 standard deviations is chosen. It was also shown as an example for the evaluation method for the EMD in figure 3.1. In figure 5.1 the result of the EMD in subfigure 5.1a is compared to the results of the KLD (subfigure 5.1b), JSD (5.1c) and RMS (5.1d). The shift of a mean of $y_{sd} = 0.4$ standard deviations is around the value that reaches the detection limit by the EMD method. By visual inspection it becomes obvious, that EMD is the only method, which shows a pattern in this case. The other three methods deliver a random result, which does not allow the detection of a shift in mean at the position given by the mark 'ls'.

Another remark has to be made on the KLD. Like it was said before, in the definition of the measure in section 3.3.2.1, the KLD is asymmetric. It was also said in section 3.3.3.1, that in this case the better result of two possibilities is used. While for the shift in mean test the difference was negligible, it was remarkable for the shift in variance. In section 3.3.3.4 it was mentioned that in this case the better result for increasing the variance from the two possible options was used. As a consequence, the results for

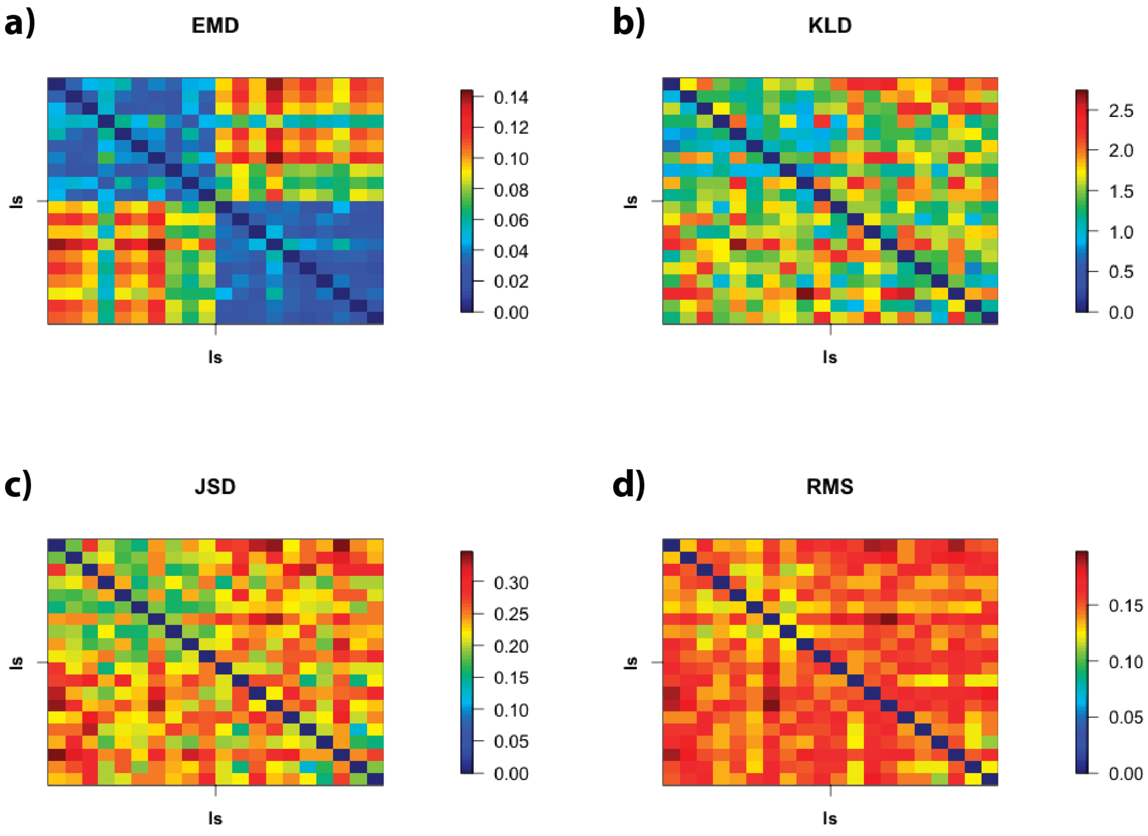


Figure 5.1: Comparison of the histogram test with the different measures on a normal distributed test vector with 2000 elements including a step of 0.4 standard deviations in the middle. In the upper left subfigure the results for the EMD, in the upper right for the KLD, in the lower left for the JSD and in the lower right for the RMS are shown.

decreasing the variance performs worse than the other option. In the other case, the KLD behaves like the JSD, RMS and MS for decreasing the variance, but performs much worse for increasing the variance of the vector under consideration.

5.1.1.2 Influence of the measures on the results

Although the EMD performs best in cases of a shift in mean or variance, this is not the case when rounding within a dataset is under investigation. This is shown in the application with the Lindenberg wind measurements in section 4.1. Here, the bin-wise comparing measures, especially the KLD, deliver the best indications for this phenomenon. The result is independent of the chosen number of bins n_b, what means, that the EMD performs worse even with a higher n_b than the other measures.

Reasons can be found in the general functionality of these measures. The bin-wise methods, which are KLD, JSD, RMS and MS, take a seperate look at every bin and calculate the difference between the two histograms for each bin. Those separate results are summed up, in order to determine the distance between the two histograms. The EMD works different. Like it was already shown in the introduction of the EMD in section 3.3.2.3 it is the solution of a transportation problem. Therefore, not only the actual

difference between the two histograms is of interest. Also the 'distance' between two bins, between which the probability has to be transported, needs to be accounted for the EMD result.
As a consequence, in situations where the differences for some bins are immense, but the needed distance between these bins to bring the two histograms to a match is low, the bin-wise methods are superior in detection. This situation is typical, if a rounding takes place within the dataset. In the other cases, like shift in mean or variance, it is the other way round. Here, the distance for the probability to transform one histogram to the other is big, but the difference for each bin is not. The EMD delivers much better results, just like it was discussed in the last section.
Additional uncertainties of the performance of a test with the KLD and JSD measure are included by the choice to use a prior information in their calibration. The option chosen here, which calibrates the method on the detection of rounding data, would of course influence the results on the detection of rounding itself. Nevertheless, the basic phenomenon explained above holds.

5.1.1.3 Reduction of information

A far-reaching consequence of the histogram test, like it is presented here, is the reduction of information. The probability density of blocks is estimated from of a defined number of values. The estimation is then compared to the other blocks. This leads to a reduction of information, which may be of advantage, but can also be a disadvantage.
The advantages have been shown in the applications in the sections 4.1 to 4.3. First of all, it allows to detect inconsistencies, which a human can hardly detect by visual inspection alone. By the application on time series of two-dimensional data it was also shown that it is possible to find inconsistencies in multidimensional datasets. Doing this, the method is only limited by the possibility of the used computer to calculate the histograms.
Disadvantages can be seen in the problem that the identification of patterns is subjective. For example, It can be influenced by the used colour table to plot the matrix. Even if patterns are found, it is not sure, if there is really an inconsistency within the dataset, since they might also be introduced by chance. Furthermore, the division of the dataset into arbitrary blocks might produce artefacts, which might be misinterpreted. Therefore, it is essential to be aware of periodic patterns within the original dataset in order to prevent their influence on the results. A possible way to achieve this is to choose the size of a block in the order of the lowest frequency of the known included periodic cycles within the dataset. This was done in the here explained application. Other possibilities can be the exclusion of annual or diurnal cycles of datasets by only looking at the anomalies of them.
To minimise the influence of the subjective evaluation, the combination with a change point model is a good solution. The method by Dose-Menzel used in this thesis, will be discussed in the following section. Other possibilities for the future may be to use the result matrices as a basis for adjacency matrices for complex networks (Strogatz [2001], Boccaletti et al. [2006]). If the characteristics like the average path length or the clustering coefficient would be sensitive on patterns, they might offer a possibility to evaluate the information of the whole matrix at once.

5.1.2 Change point model

The detection of breakpoints in time series is an important tool in quality control and assurance of datasets. In this thesis, especially the method by Dose and Menzel [2004] is investigated and modified. Apart from using this method in its basic configuration, it is used in combination with other methods like for example the histogram test. Its special advantage is a result, which is given as a probability and which may be usable in a framework like the quality evaluation. This advantage makes it an important tool to integrate quality tests into such systems by translating their results to a usable form. First, the Dose-Menzel method and its usability in the environment of quality assurance will be discussed in section 5.1.2.1. The main topic is the analysis of the sensitivity tests in section 3.4.3. Further discussion follows in section 5.1.2.2 with the focus on the problematic of interpretation of the results by the method of Dose and Menzel. Also, the method to detect multiple change points is briefly discussed here. In the following section 5.1.2.3 the inter-comparison test with other detection methods for change points performed in section 3.4.4 is discussed. At the end of this section, the combination with other methods like the histogram test will be addressed in section 5.1.2.4.

5.1.2.1 Dose Menzel in quality control

The method by Dose and Menzel, like it is introduced in Dose and Menzel [2004], was mainly used for applications of data analysis in phenology. Applications in quality control or assurance are unknown up to this date. In section 3.4.2.3, two further main modifications are introduced besides the original modification of this data analysis tool. In the following section 3.4.3, some sensitivity tests are performed to show the possibilities and limits of the method. In these sensitivity tests, the parameter γ is identified as a critical parameter, what allows to modify the sensitivity to detect a breakpoint of the test. This becomes especially obvious in section 3.4.3.1, where the three models, constant, linear and one change point, are compared for different γ for all three main modifications. The difference of the dependence of the three modifications on γ is not immense, but does exist. This was explained in section 3.4.3.3, where it was shown that different γ are needed to reach the same detection limits. Generally, the flat model is more sensitive for lower shifts in the mean of a vector than the other two modifications for the same γ. The same can be seen in section 3.4.3.4, where the modifications are calibrated to have a false alarm rate of less than 5%. Using homogeneous vectors and the false alarm rate to calibrate the model, like it is done there, was also performed by Ducré-Robitaille et al. [2003].

Bigger differences between the three modifications could be found through the investigation of the detection limit under the condition of the position of the step in section 3.4.3.2. Here, the same γ was used for every method, but especially the original method showed its limitations to detect steps at the end of the vector.

An advantage of the method by Dose and Menzel is the possibility of flexible models, which are regressed to the data. Only linear regression models of different styles are used in this thesis. Other possibilities for models are waves with different wavelengths and phases. Those may be further options for using the method in data analysis. Nevertheless, in a quality control environment it is appropriate to exclude known periodical cycles and use calibrated linear regression models, like it is done in this thesis.

5.1.2.2 The abilities of the Dose-Menzel method

Detecting change points is always a complicate application of statistical methods. It is hard to decide, whether a detected change point is really a change in the measurement environment or if it happens just by chance. The Dose-Menzel method uses an approach, which makes the decision process more transparent. By comparing models to each other it is much simpler for a user to decide in doubtful situations. Theoretically, he/she has the option to include further models and compare all of them in the same framework. The method is also applicable under the condition of automatisation. These attributes make it also a possible tool for homogenisation applications, since it would allow the introduction of estimations of uncertainties for detected inconsistencies.

In case of multiple change point detection, several approaches are possible. They include for example the direct implementation within the model equation, what means that the whole vector with models of different positions and different number of change points are checked. Problematic is here, that the

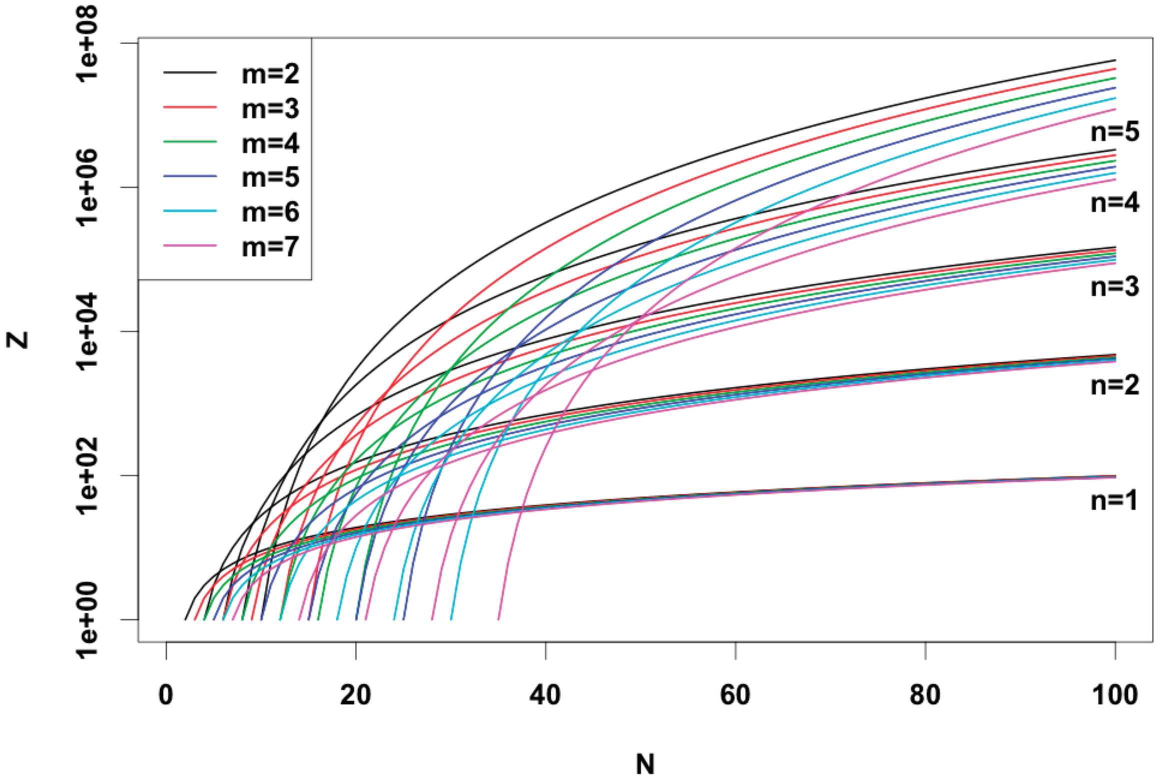


Figure 5.2: Number of possible combinations (Z) for a given number of elements (N), number of breaks (n) and the minimum number of elements in a linear section (m). The x-axis shows the length of the dataset under consideration N and the y-axis the number of the combinations Z on a logarithmic scale. Colours indicate a different choice of m, the different branches are generated by the different number of assumed breakpoints n.

number of possibilities of change point settings Z, rise dramatically for further included change points. Some calculations of algebra lead to:

$$Z(N, n, m) = \binom{N - n(m-1)}{n} = \frac{(N - n(m-1))!}{n!(N - n(m-1) - n)!}. \tag{5.1}$$

Here, n is the number of change points, m the minimal length of a linear section and N the length of the vector under control. The equation 5.1 holds under the assumption, that it is sufficient to check the minimal length of a linear subsection. If not, the sum of $Z(N, n, m)$ from the minimal m to the maximal possible length of a linear section, which is given by the length of the dataset N, has to be taken. The equation is only valid for a defined number of change points. Should several different numbers of change points be checked, the results would have to be summarised between the minimum and maximum number of assumed change points. The raise is shown in figure 5.2. The number of combinations Z is shown here for different N, n and m. Obviously, the order of possible combinations depends mainly on the number of elements of the vector N and the number of assumed change points n. It shows, that the number of necessary tests, to check an immense amount of change point combinations for longer time series and several change points, would be very high. This would reduce the possible application of this approach to short time series.

Another possibility to detect multiple change points with the Dose-Menzel method is the here used recursive method. It divides the dataset at the most probable change point and performs, if possible, further analysis of the two resulting subsamples. The advantage is, that it can be programmed recursively, what makes this way of multiple detection effective. A problem is that a breakpoint in a subsample does not automatically have to be a breakpoint for the whole time series. The method was used by Ducré-Robitaille et al. [2003], but is not state of the art anymore. Nevertheless, as this thesis uses comparisons with the cited article, it is used here. Other approaches like the dynamic programming, which was introduced into homogenisation by Hawkins [2001], are more common nowadays. Especially, if change point models are used in homogenisation the procedure delivers good results (Venema et al. [2012]). For applications like the quality evaluation procedure it is reasonable to use only the information on whether the time series under investigation is inconsistent or not. The number of change points and their position is only of interest in the further analysis on where the problems for the quality within the dataset occur.

5.1.2.3 Comparison to other tests

In section 3.4.4, the main modifications of Dose and Menzel are compared to other change point detection methods gathered from Ducré-Robitaille et al. [2003]. The used test settings are taken from this article as well. Modifications were made through the use of additional types of test vectors and the type of evaluation of the performed tests.

The implementation of most of the methods by Ducré-Robitaille et al. [2003] seems to deliver comparable results to the ones found in the cited article. This has been shown in section 3.4.4.3, where a test on homogeneous time series was performed. The methods, which apparently could not be reprogrammed properly, are the MLR and the Bayes with reference method. Concerning the MLR, the reasons may be found in the usage of the R package (Pinheiro et al. [2011]) for some calculations. Here, several problems occurred, which might leave artefacts in the results. The problems within the Bayes with reference series method can mainly be attributed to a resolution problem. In this method, an integral has to be discretised and evaluated. The originally used discretisation in Ducré-Robitaille et al. [2003] is unknown, what may lead to some uncertainties in the results. This is problematic, because the same function was also used for the Bayes without reference series, which produces valuable results. Therefore, this resolution problem is especially emphasised, when the variance of the controlled time series is small. This is for example the case, when a reference series is subtracted from an original series, what is done for the Bayes with reference method. Another indication for this behaviour can be seen in the differences in the sensitivity of the Bayes with reference method to the different types of test vectors. By construction, the variance of a difference series for the autoregressive test vectors is higher than for the normal and gamma distributed ones. As a consequence, this leads to a lower sensitivity for the method on the autoregressive test vectors.

For all performed tests in section 3.4.4.4, that include inhomogeneous vectors, the modifications of Dose and Menzel using the reference series show better results than the modifications, which do not. Due to their calibration, which was performed in section 3.4.3.4, the other parameters like the threshold and the type of the model do not have a lot of influence.

With the reference series that is also used by all but one method described by Ducré-Robitaille et al. [2003], the method of Dose-Menzel shows similar results as the two modifications of the SNHT. These methods are also the ones that performed best in all types of tests and analyses carried out here. The other methods do all have deficits in each analysis, what can partly be traced back to the mis-calibration, shown in the test on homogeneous datasets in section 3.4.4.3. In addition, unlike the analysis of Ducré-Robitaille et al. [2003], the investigation here focuses only on the ability to detect an. The dependence of the detection rate of all methods on the size of the steps is not analysed in this thesis.

5.1.2.4 Combination with other methods

The combination of the Dose and Menzel method with other quality checks has been shown within the inter-comparison study in section 3.5 and as an application in the sections 4.1, 4.2.1 and 4.2.2. Especially the combination with the histogram test was demonstrated here. That it is also possible to combine Dose and Menzel with other methods, like to the distribution tests presented in section 3.2, was shown in an application in section 4.2.

The inter-comparison tests showed that the application of Dose and Menzel to the results of the histogram test with the EMD was very successful. Even when the size of a block s_b is low, the results for the homogeneous vector are comparable to the application of Dose and Menzel on the raw datasets. For the inhomogeneous datasets, a lower sensitivity was detected when the histogram test was used as an intermediate step. This might be corrected by recalibrating the combined method as a whole and not only the part of the change point detection.

In the applications in chapter 4, this approach performed well in detecting inhomogeneities even in multidimensional time series. Nevertheless, the positioning of steps highly depends on the used regression model. It has been shown, that detected inhomogeneities can be traced back to changes in the generation of the datasets like changes in the types and numbers of instruments, of the assimilated observations for the reanalyses. The sensitivity of the different measures for the types of inconsistencies is identical with the one that was already described in section 5.1.1.2. While the EMD is more sensitive to steps, the KLD and JSD are more sensitive to rounding within datasets.

Should the dataset under consideration have several inconsistencies, the combined method with the histogram test might miss them. Furthermore, if their effects neutralise themselves within the regression, it might be possible that no inconsistency is detected at all. This has to be kept in mind, when the method is used under conditions of automatisation. If a dataset is classified as inhomogeneous, the results indicate, that this is indeed the case. For datasets classified as homogeneous this is not necessarily the case. The same phenomenon was shown in the inter-comparison tests above.

An advantage of the procedure is the generation of probability information except a basic method, that originally did not deliver one. Apart from the histogram test it can be applied to any one dimensional time series. This time series might be the result of another quality check, like it was shown with the block-wise calculation of statistical parameters. The latter can be used for example within quality evaluation, which will be discussed in the next section.

Furthermore, the usage within homogenisation procedures of the combined and the uncombined method of Dose and Menzel is possible. Since the combined method allows to analyse multidimensional time series with respect to inhomogeneities, it might provide new approaches for their homogenisation. The results, which are given as probabilities, would also allow to estimate uncertainties of the corrected time series within a homogenisation process.

5.2 Quality evaluation

In section 2.6.7, quality evaluation was introduced as a simplification to evaluate the SQA on data. It is based on the assumption, that more than one quality check is necessary, in order to determine the quality of a dataset. The aim is to perform an estimation of quality with the knowledge of an expert. Additionally, the system has to be able to effectively evaluate immense amounts of data and tests with the intention that the time of a quality controller is not unnecessarily wasted. It also has to be a transparent process and should uniquely quantify the quality.

This section starts with some general remarks on the used process in section 5.2.1. It focuses on the general possibility to evaluate the quality of data with the help of automatisation techniques. Afterwards, the advantages and disadvantages of the presented system are discussed in section 5.2.2. A special disadvantage, the possibility of over-interpreting the results, will be examined in detail in the subsequent section 5.2.3.

5.2.1 Statistical quality evaluation

It was already mentioned in section 2.6.7, that using statistical tools to estimate the quality of an observation is nothing unusual. An example is the data assimilation within numerical weather prediction models. The calculation of probabilities of observations on basis of the background of an existing forecast was performed by Lorenc and Hammon [1988] and Ingleby and Lorenc [1993]. Quality assurance of data traditionally means that every datum is checked separately and flagged, if the test indicates that it is suspicious (Hubbard et al. [2005]). Alternatively to the flagging of values, the ECMWF introduced percentages (Baker [1992]). The view of separately evaluating the quality of every datum was modified in this thesis, since the aim is to obtain a quality information on the whole dataset. Therefore, automatisation is necessary, as in case of large datasets it is impossible to check every data value by hand (Gronell and Wijffels [2008]).

Quality evaluation in the here presented form tries to represent the quality of a dataset in a standardised form and with only one value at the end. Within this framework, the parameter Q, which is used in section 2.6.7 as a measure for good data quality, can be given a different meaning when needed. A representation as bad data quality for example would not be too complicate to implement. The changes to be performed are the used prior information and the calculation or interpretation of the probability displayed by the tests.

The used model is not the simplest possible model to solve the assigned task. One difficulty is for example the inclusion of the general prior out of technical requirements. This prior requires the expert to specify an equivalent for the quality of the dataset for every possible result of the quality check . This is achieved by a function, which by convention should include at least one result, where the quality estimate is maximal and therefore equals 1. A possible simplification, that could be used here, is that the result of the check directly defines also the estimate of the quality. Since it is defined as a value between 0 and 1, there would not arise any technical problems. In this case, it would lead to a simple linear function for the general prior in the method used here. Generally, it cannot be expected, that the relationship of the results of a quality check is linearly connected to the quality estimation. An example is an outlier test. If there are just a few outliers, that contain only a small rate of values of the whole dataset, this might be acceptable for a dataset. A linear relationship would lead to the consequence, that if half of the values are outliers, the quality is at 50%. This is not realistic for a lot of physical variables, because many outliers might indicate an unusable dataset, if the limits are defined as physical constraints of the measurements. Therefore, the results of the quality estimation would be expected to be low in this case. Of course this expectation could be directly considered in the calculation of the percentages of the quality check. Still, it will lead to more transparency, if the knowledge of the expert is disconnected from the test itself. Additionally, the fact that the expert has to give the prior in advance, is a technical requirement, which can be abolished. It might be possible, that the expert is asked from case to case, how he/she would assign a quality estimation to a given test result. Nevertheless, this is not acceptable in cases where automatisation is a basic requirement, like for larger datasets.

The additional inclusion of a weighting is also arguable. A simple example for the need of a weighting can be given with the LIM-test as well. In the case of a temperature measurement in summer in Germany, a possible set of limits can be defined by 10 to 35 degrees Celsius. Of course, the temperatures might exceed those limits. A useful determination of the limits is given by the guidelines for the quality control of automatic weather stations by the WMO (Zahumensky [2007]). It sets the acceptable range to extreme values (-80 to +60 degrees Celsius). If those limits were exceeded, this would really be an acceptable outlier. Nevertheless, the information of the narrow range can be of help for a quality estimation as well. In this case, both tests can be performed, wherein the wide range is checked with a high weighting and the narrow range with a lower one. Consequently, both pieces of information are adequately represented in the quality estimation.

An arising question is the necessary amount of effort needed from the user. It depends, of course, both on the amount and type of data to be quality controlled. To quantify this, the effort of a user has to be divided into two parts. The first is the active work, that a user has to perform himself/herself. As a second part, the work is defined, which he/she is able to externalise to a computer. The amount of work a user has to perform himself/herself, mainly depends on the number of types of data within the dataset under consideration. That is due to the fact, that for every type of data, the tests, parameters, priors and weights have to be defined separately. After the manual pre-definition, the automatisation of the quality evaluation algorithm allows that a computer is able to perform the rest of the necessary work to reach an estimation of the quality. For the user this is again independent of the amount of data to be checked. It allows a user to only check the quality estimations at the end, when the algorithms are implemented and the knowledge of the expert is put into place.

From the technical point of view it would be simple to include quality evaluation into a software system like 'qat', which was presented in section 2.6.3. There, it could be implemented as a fourth branch besides the analysis, the plotting and the saving part. Nevertheless, there are parts that have to be developed anew, more precisely an exchange formate for the prior and the weighting information for the tests. A possible solution might be the development of an XML scheme, just like for the tests and parameters.

In a scientific quality assurance on data it is expected, that the quality estimation is followed by a careful examination of the suspicious data through the author. Quality estimation only simplifies the detection, but cannot compensate for the interpretation within a scientific quality assurance on data. In the example given in section 4.4, an expert would have the task to connect the findings of the quality checks to the physical reasons. If this was performed graphically, the expert would have to evaluate more than 6300 quality checks, with one or more graphics as a result. The quality evaluation mechanism on the contrary enables him/her to connect the datasets, for example with the explanation of the range of the wind direction measurements, in a simple and effective way. This assists a data reuser in using the datasets on his/her own.

5.2.2 Advantages and disadvantages of quality evaluation

This section further investigates the advantages and disadvantages of the presented quality evaluation system in section 2.6.7. It starts with a look at the general concept of the process and afterwards goes into more detail concerning the components.

The idea to evaluate all kinds of tests, to be performed on one dataset on the same basis, is generally an advantage. With this, it becomes possible to use just one framework for the evaluation of all tests. A disadvantage is, that the tests are required to deliver their results in terms of probability. Since most tests do not have that ability, additional steps are needed to transform the results to a usable format for this type of quality evaluation. A very helpful tool is the change point detection by Dose and Menzel, which was presented in section 3.4.2.1. It allows to check time series on whether they include inconsistencies or not. Of course, those do not have to be the original time series of the dataset under consideration. A possible step in between, can be for example the determination of statistical parameters in section 3.2.2 or the histogram test thar was shown in section 3.5. An arising problem is to establish an acceptable documentation. The test is no longer only driven by the parameters of one test, but of two. The additional complexity might irritate the data reuser.

Another advantage is the flexibility of the described process. It allows to test sets of parameters, which are otherwise not usable. The outlier check on temperatures in Germany in the last section 5.2.1 was given as an example. An important tool are the priors and weighting factors. One problem is the definition of a general prior for a test. An expert does not only have to determine the prior information for the most common outputs of a test, but for all. Additional complexity arising through intermediate steps can make this a complicate task. Therefore, it is important to determine a clear null hypothesis on whose basis the probability result of the test is calculated.

Priors itself are also not without risks. An expert does not only have to know the data and the expectations on the data, but also the behaviour of the used quality checks that depend on the used test parameters. It is also expectable that different experts would use completely different priors. This is an advantage and a disadvantage at once. On the one hand it makes the system very flexible, since different users are able to set their own standards to determine their own quality estimation. It might be possible to define three different standard situations in which this process can be used as a practical application: The first is the usage of the data author himself/herself. He/She is able to analyse, if the data are like he/she expected them to be. The second is the data reuser, who wants to recheck the data with his/her own parameters or additional tests. In this context he/she is able to use his/her own priors. As a third type of users, reviewers of an SQA on data can be assumed. They are enabled to use different priors, different weightings, different parameters or additional tests to verify the statement of the data author on the quality of the data. This third possibility will be of interest in section 5.3.4.2, where a peer review of data will be discussed.

An additional problem is that the result of the quality evaluation is only relative. Without a standardisation of the underlying tests, parameters, priors and weighting factors the quality estimation is only an usable information for those, who know the circumstances of their definition. There is no definition, what a quality estimation of for example 65 % means for the user of the dataset. The information is only helpful, if the quality is compared to similar datasets in the same test environment.

The last remark leads to a strong disadvantage of the system: the possible over-interpretation of the results, that will be discussed in the next section 5.2.3.

5.2.3 Risk to over-interpret the results

A possible risk originating from quality evaluation is that if someone gives a quality estimation of a dataset, others will take it as a definitive statement. They might believe in it and use it in their own work. So what would happen, if a dataset earned a quality estimate of 100%? Is it really a top dataset without any problems? It depends on the definitions. Like described above, the results of a quality check depend highly on the selected tests, parameters, priors and weightings. Additionally, it is possible that data considered perfect for some people, might be imperfect for others. This is a valid threat to an absolute statement (Parsons et al. [2010]). For example scientists, who want to introduce the data to run models do not like it, if there are possible outliers in the dataset, which might irritate their own model. On the other hand, the ones who are working in the field of extreme value theory are extremely interested in those values. For whom is which dataset a perfect one? This question cannot be answered and should not be the problem of the method itself. Quality estimates, especially when their output is only only one value as a result, have to be used with extreme caution by all parties involved. Therefore, it might be imaginable to minimise even this type of information and transform the resulting numbers

for example to a traffic light system. Green would indicate a good dataset, yellow some doubts and red would warn against a highly dubious dataset. Nevertheless, only one information cannot represent the quality of a whole dataset, since the requirements of the data reusers are too different.
If different interest groups for important datasets exist, it might be useful to generate different profiles for different target groups. This enlarges the effort and leads to even more pressure to standardise the test environments for typical users of data. Another possibility to show that results of a quality evaluation are only relative, would be the inclusion of uncertainties. The repetition of the tests with different parameters is not recommended for larger datasets, since this might be very computer intensive work. It would be simpler, to use bootstrapping algorithms, where the priors and/or weightings are varied. The influence of the expert choices of these subjective parameters on the quality estimation can thereby be illustrated. Nevertheless, whether this quality estimation is still usable for a normal scientist, is more than doubtful.

5.3 Data publication

The importance of the publication process in science today was extensively described in chapter 2. The primary data publication was characterised as an evolving kind of a scientific publication introduced in section 2.6. Its embedding into the scientific process is illustrated by the scientific working scheme, which will be discussed in section 5.3.1. The scheme shows, that it fits into the traditional work of a scientist and that it defines the requirements of new developments of quality assurance procedures to be comparable to the other forms of publication. Another important attribute of the publication process is its transparency for all involved parties in section 5.3.2. It will be shown that only if such a process is transparent it can be accepted in the scientific daily work. Afterwards, some remarks will be given on the possible ways to introduce the concepts into a data centre in section 5.3.3. The practical steps and the use of these more theoretical concepts will be illustrated. In the last section before the conclusion of this thesis in section 5.3.4, it will be commented on how the concepts can be used to introduce an effective data peer review in the future.

5.3.1 The scientific working scheme

The scientific working scheme to be discussed in this section, was introduced in section 2.2.2. Further enhancements were included and discussed in section 2.3.4. In the following, some general remarks on the working scheme will be given in section 5.3.1.1. Since the application of this scheme in science is plausible because of its flexibility, it is also discussed here. The scheme allows to introduce quality assurance steps for the different types of publication. A further discussion of these steps follows in section 5.3.1.2. Finally, the difference between the terms 'quality assurance' and 'quality control' will be examined in section 5.3.1.3. The intention is to show, how the term 'quality evaluation', which is used in section 2.6.7 and discussed in section 5.2, fits into this context.

5.3.1.1 The working scheme in general

The main purpose of the scientific working scheme was to show, how new forms of scientific publications, namely data papers and primary data publication, integrate into the work of scientists. Therefore, a possible general working scheme of a scientist, which is similar to a standard structure of a scientific paper, is set up. It is used to determine, which part of the work of a scientist has to be introduced into which type of publication. Since all working steps are introduced into a traditional paper publication, the latter can be seen as the most important form. The two others only cover parts of the working steps, but therefore allow to go into more detail. While the data paper focuses especially on the experimental design and the experiment itself, it offers a very good opportunity for a reuser of the performed scientific work to get introduced into the measurements. Therefore, data reusers are enabled to decide, whether the dataset is useful for them or not. The primary data publication, with its two parts of data and metadata, allows to access the base of a research by others. Like it was explained in section 2.3.2 this is not only regarded as an opportunity, but also as a threat. In addition, it is not enough to simply make the data available in the world wide web. Rather a lot of effort is needed in order to make it usable for the scientific community. One elementary part is the quality assurance process and its proper documentation.

This thesis described in several steps, how these quality assurance steps can be constructed. Further steps, like the peer review, are established in science for a long time, but are not accepted for primary data publications yet.

A special feature of this scheme is its high flexibility. Since not all mentioned elements of the scheme have to be performed by a scientist, it is still possible to use. The scheme shows, which parts should be included in which form of publication in order to obtain an effective result for the data reuser as well as for the author of the experiment. Apart from field experiments it is usable for data classes like model data. In case of such data, modifications of the documentation steps have to be done. Since the model codes, input fields and the used parameters are a crucial part of the performed experiments, these data have to be well documented and quality assured as well. The quality assurance steps can be modified appropriately, if it is necessary. In the next section follows a discussion of the quality assurance steps explained in this thesis.

5.3.1.2 Quality Assurance processes

In section 2.4.2, three different quality assurance processes were purposed, which will be discussed in this section. They are constructed for three different types of data: the quality assurance on methods for free form text, the quality assurance on metadata for standardised metadata and the quality assurance on data for the unstructured data itself. All three are only proposals and can be varied, if required. Nevertheless, they cover the actual developments well.

The quality assurance on methods works on the existing traditional and data paper. Both are free form texts and therefore need special attention. Since it depends on the journal, which part of this quality assurance will be done in the scope of the author and which by the staff or the technical guidance of the journal, the proposal has to be interpreted as a flexible one here. Traditionally, the collection of information and writing of a manuscript is done by the author, as part of his/her analysis. Since the analysis itself covers the investigation of the data and the experiment backed by the theory, the process can also be separated. The steps needed for the quality assurance are only the technical parts. A submission to the journal can either be done before or after the discussion part. Discussion papers, like they are described in section 2.2.3, are generated after the submission, while pre-publications can be created and distributed before. In addition, the transfer from the quality assurance to the peer review process might be not as strict as defined here. Therefore, it is possible to include additional quality measures, like spell checking and proof reading, even if they are performed after the peer review process. These measures would not be a quality assurance in its original purpose, but a quality control. The difference between these two terms will be explained in the next section 5.3.1.3.

The quality assurance of metadata is indeed a mix of a quality assurance and a quality control. The collection of the metadata can happen in several ways. Technical solutions, like their extraction from the original data files, are a basis, that can be simplified by the use of self-describing data formats like netCDF (Talbott et al. [2006], Eaton et al. [2011]). Tools, that were developed for the completion of the metadata, like Atarrabi, can also be used to perform their initial collection. This software was originally designed in the project "Publikation Umweltdaten" to work on field campaign data. With some modifications it is also usable for model data, like the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Quadt et al. [2012]). Further flexibility would be needed, should it become applicable for different requirements of metadata. The quality measures implemented within Atarrabi are useful to enrich the quality of the dataset (see also section 2.5.4).

The quality assurance on data has to be divided into several steps. It starts with the quality assurance in the field, when the data is collected. This generally includes quality control procedures, that aim at the correction of false datasets (Meek and Hatfield [1994]). The quality assurance performed within the publication process itself, should not change the data anymore. It should rather describe the performed measures in order to help the reuser to handle the limitations and problematics of the dataset. Therefore, an appropriate documentation of the quality assurance is essential and has to be set up as standardised as possible. The information is not directly stored within the dataset, but in the metadata. The latter is another quality measure. It ensures, that the dataset is not changed anymore after it has been submitted to the data centre.

The different quality assurances share a common basis. Their task is to increase the quality and to help the reuser to use the published information. In the next section, the difference between quality assurance and control is further explained, since both are similar, but not synonymous.

5.3.1.3 Quality assurance or quality control

In literature, there are mainly two terms used for processes, that help to ensure the quality of an entity: quality assurance and quality control. Therein, a lot of different definitions are used for the two terms sometimes contradicting themselves. The World Meteorological Organization (WMO) gives a definition in context of automatic weather stations in Zahumensky [2007]. Quality control is defined as: "The operational techniques and activities that are used to fulfil requirements for quality." This also includes "missing data detection, error detection and possible error corrections." In contrast, the definition for a quality assurance is given as: "All the planned and systematic activities implemented within the quality system, and demonstrated as needed, to provide adequate confidence that an entity will fulfil requirements for quality." Its aim "[...] is to ensure that data are consistent, meet the data quality objectives and are supported by comprehensive description of methodology".

At first glance, both definitions seem to be quite similar. Still, there are important differences. One possible solution in a quality control system, in case of an existing error, is to correct it. On the other hand, a quality assurance system just documents the error under these definitions. Another difference is, that a quality assurance system aims at making data consistent, while a quality control system tries to reach the best possible result for an entity.

Going back to the procedures for the three different kinds of quality assurance in section 2.4.2, it is hard to decide, whether all three types are purely a quality assurance under the definition given above. In the case of the SQA on methods, which is used for free text documentation in papers and data papers, it is definitely aimed to be a pure quality assurance. Discussion papers or pre-publications, which form the controlling part of this SQA, do not aim at the correction of the text under control. They are thought to give additional input to the publishing scientist. If he/she decides to correct parts of the publication in a next step, this can mainly be seen as a step in the peer review process. There, it is obviously an important aim to obtain the best possible result for the publication. The confidence on the quality can for example be enhanced in a discussion paper. There, the publishing scientist can explain arising questions to prevent misunderstandings that might arise from his/her text.

In case of the SQA on metadata the process is different, because it shows mixed characteristics. Its aim is more a quality control, since it is hoped that the publishing scientist corrects and completes the erroneous or missing metadata. Furthermore, the validation part is a quality control procedure, since it forces the data author to correct wrong entered information. The characteristic of quality assurance in the SQA on metadata, can be seen in the aim to make the metadata consistent. If metadata is imported from a different metadata convention, it dose not necessarily have to be an error, if a certain metadata entry is missing. The SQA on metadata therefore aims at meeting the quality objectives of the actual system. It does also not force the user to fill out all possible metadata entries. The required metadata entries are just a small necessary subset of all possible metadata. Therefore, the term SQA on metadata is used throughout this thesis, even though it is partly a quality control system.

The third kind of quality procedure, the SQA on data, is a strict quality assurance. It controls the data, but does not perform any corrections. The aim is to document and comment on suspicious results as well as possible.

The new introduced term of quality evaluation is not directly connected to the above discussed two terms. It rates the quality of the dataset under given constraints. Therefore, it might be used within both of the procedures of quality control and assurance in order to get an overview over the performed measurements. Of course, this can be extended to the SQA on metadata and methods, if it is possible to estimate a good quality in these procedures by tests. For the metadata, this could be tests on the number of given metadata entries or the result of a validation. Nevertheless, it seems to be only of practical use for the data part.

5.3.2 Transparency as a basic foundation of data publication

After the explanation of the technical background of data publication, another foundation will be enlighted in this section. Transparency within the whole procedure is a basic requirement needed to enable other scientists to make use of the datasets for their own work. Some measures, which can be helpful to grant transparency, will be discussed here. The necessities of the measures to make data publication a success, are strongly connected with the generation of trust in datasets. As it is the key to generate citations and therefore credit for the author, transparency is a basic requirement for a successful publication system.

Like it was already said in section 2.3.1, trust into the data is required, if others are to use it and not to re-perform everything on their own. The best way to generate trust is transparency. It is not only a must for the performed quality assurance, but also for the whole editorial process. The latter was characterised by Costello [2009] as a part of "scientific editorial standards". Without explicit measures to guarantee these transparency it might easily be compromised.

One transparency measure is based on the fact, that a data author has to know what happens with his/her data if he/she forwards them to the publisher. This is ensured for example in the metadata completing software Atarrabi by the extensive help texts mentioned in 2.5.4. Additionally, the opportunity to communicate with the publication agent within the software might be helpful. Both measures are obviously part of a transparent editorial process.

The transparency measures within the quality assurance steps can be found in a detailed documentation, which allows a data reuser to reconstruct the performed quality checks (see also 2.6.2). That this is important was stressed by Durre et al. [2008], who reasoned that undocumented quality control and assurance can compromise the interpretation of the datasets.

Still, transparency is not the only measure needed to build up trust in the published datasets. Further possibilities to make the datasets trustable will be discussed in the following section.

5.3.3 Integration of quality assurance at data centres

This section will illustrate how the described methods can be included at a data centre like the WDC-C. The inclusion is based on the technical foundation that was laid in the project "Publikation Umweltdaten" and especially on Atarrabi. The section starts with some basic discussion on the documentation of quality checks and the needed accuracy in section 5.3.3.1. It will be described, how the documentation can be used as an interface, to take further steps on the inclusion of a scientific quality assurance on data in a data centre. One of these steps is the calculation necessary for the quality assurance on data, at a data centre itself. The necessity and the problems will be discussed in section 5.3.3.2. This includes for example the necessity of standardised tests. These developments would be a basic requirement for a data publication on the same level accepted as traditional publication. They would also lay the foundation for data peer review, which will be discussed afterwards in section 5.3.4.

5.3.3.1 Documentation of quality assurance processes

The software Atarrabi that was presented in section 2.5.4, uses a different view for every topic of the documentation of the datasets. This documentation is done by the modification of their metadata. One view, that was shown in figure 2.7, is used to document the performed steps of the quality assurance on data. Currently, it focuses on the data level (see also section 2.6.2) and the approval and allows both a comment on the quality and the upload of files. The used granulation is twofold, since either this information can be given for the whole experiment or for every dataset separately.

The view of the actual version (2.1) of Atarrabi is the starting point for the documentation and allows, due to its flexibility of the file upload, a high grade of documentation with a low amount of effort for the data centre and the user. The question arises, which granulation and grade of accuracy would be optimal for data publication and its quality assurance steps. The answer is complicated, since it depends on the roles played by the scientists. For the user, who published the dataset, the grade of accuracy should be low, since it is a lot of effort to fill out very detailed forms. For the data centre it is also simpler if a low grade of accuracy is used, because it leads to less needed storage and a smaller amount of maintenance due to simpler database structures. Nevertheless, if it is only achieved by making use of file uploads to compensate for the low grade of accuracy, the storage argument might be compromised. The third party involved, the data reuser, might be interested in a high grade of detail, as long as it is well presented to him/her. Immense amounts of data are hard to access by a user, if he/she wants to get the correct information he/she needs. Nevertheless, it can be necessary for him/her to have the specific information of the data, with a high granulation, if problems arise. In addition, it would be desirable to have as much information as possible, under the condition of a simple accessibility. This is especially reasonable in terms of transparency, what was discussed in the last section 5.3.2.

Taking the position of 'the more information the better' would have consequences apart from the possible workload. Those consequences will be discussed in the next section. In preperation, a possible grade of accuracy will be explained here. Ideally, the data is documented on the level of every variable in every dataset of the experiment. This would mean, that the information can be reduced to one measurement vector or field, when the basic information is a time series. On that level, the documentation can be done test-wise, like it was explained in section 2.6.2. The name of the test, a description, information on the used algorithm and parameters and a comment on the results have to be documented.

Under such a configuration it would be very important to use technical measures in order to minimise the workload for the publishing user. For example, if hierarchies are available within the dataset of the experiment. It would be helpful, if entered information, is automatically included on the lower or higher levels, if it is requested by the data author. This would need a concept of support by the documentation software through design and usability. In addition, interfaces between quality assurance software and the documentation panel can reduce unnecessary redundancies. Otherwise, they would lead to the situation, that a user has to give the information for the performed quality assurance more than once. The redundancies would even be more reduced, if standard tests for standard variables were available. A possible workflow of a quality assurance could be a centrally provided workflow for a specific variable. After having chosen the workflow, the publishing author can modify it for his/her own needs. Then, he/she can use a quality assurance software, which is able to process this workflow and enriches it with information for example about the algorithm of the test. The author checks the results and comments on them wherever necessary. This enriched workflow is then uploaded to a software like Atarrabi, which completes the metadata. Here, the workflow is evaluated and the information pre-fills the forms for the publishing user. At this point, he/she can make, if necessary, some additional corrections and then submit the information for the data publication.

That this workflow can be simplified even more will be discussed in the following section. Therefore, the tests are not performed on the resources of the publishing author, but within the data centre, which stores the data. It will also be shown, that this approach would have even more advantages.

The presented measures do all require high effort and can cause problems in means of scalability. Therefore, technical support is essential. Nevertheless, even with the above described workflow and under the assumption that standardisation works in an acceptable way, the efforts for a data publisher might be ruled as too high. In addition, a complex publication procedure might not be accepted within the scientific community, since it jeopardises the principle of transparency. Therefore, a good representation of a large amount of the data that are generated within a quality assurance, process, might be inevitable. The pivotal question is, whether the credibility, a publisher earns by the performance of a detailed quality assurance outweighs the effort. This was already argued in section 2.2.4. The experience that scientists weight the gain of credibility very high might offer ways to reach the aim that a detailed documentation will be performed by at least some scientists. As a consequence, it might be reasonable to offer different grades of granularity and accuracy for documentation to a publishing scientist, to give him/her the opportunity to choose.

5.3.3.2 Necessity of central calculation of quality tests

Some questions arise concerning the concept of data publication. For example, who should perform them and where should the calculations of the quality checks for data publication take place? The procedures presented in this thesis are designed for the calculation of the statistical tests on the resources of the publishing data author. This was already mentioned in section 2.6.3. Still, there are good reasons for the establishment of centralised calculations of these tests at a neutral place. First of all, it is the transparency which is guaranteed by the central calculation. This is much easier achieved by generating a controlled and regulated documentation of the tests. Of course, this might become possible through the usage of standardised software, but the inclusion within the workflow of the data publication is much easier.

It is also possible to minimise the risk of fraud, which might happen with the intention to obtain better results from the tests. If the success of getting a citation depended on the quality of the dataset determined by a quality assurance, there might be the possibility that some scientists try to manipulate the results. This would of course be simpler, if the software to perform the quality checks ran on their own resources. By pressing for detailed documentation of the processes, the checks can still be replicated by everyone, who has doubts on the results. The risk to get accused of fraud might be enough to limit these problems.

Another advantage is that the datasets are already at the data centre. The latter might be the required independent place, where the calculations can be performed. It is able guarantees that the controlled datasets are identical to those published later on. If a user performs the quality assurance on his/her own resources, the datasets can be compromised by transmission problems, so that the tested and the published datasets might be different.

The last argument for the calculation at a data centre is the simplicity for the user. If a user just has to give the tests and parameters and evaluates the results, he/she might be more motivated to carry out a detailed quality assurance. The steps that could be saved, are looking for the correct datasets, modifying the control software to the actual working environment, down- and uploading the results, downloading and installing the software and getting accustomed with different software environments.

But what are the requirements for doing it effectively? One possible answer is the detailed documentation discussed in the last section. The documentation of the tests, out of the comments on the results, can be done before the tests are calculated. This includes the definition of which tests should be performed on which variable and which parameters should be used in these tests. Based on this information, the tests can then be performed directly at the data centre and the results can be presented to the publishing user. He/She should then comment on the results.

The software packages, which run effectively on the infrastructure of the data centre, still have to be developed. They might orientate on the scheme, which was presented for 'qat' in section 2.6.3. Even if the hardware resources needed to calculate these tests are available at the data centre, the calculation itself might be the biggest problem for the realisation. Also, the inclusion into the workflow of the data centre has to be solved. One point here is the question where the tests can be performed, should the data be transferred out of the long term storage to the faster working storages. This has to be planned very carefully. The additional workload for the scientific staff at the data centres for maintaining the process has to be low as well.

All in all, the calculations of the quality checks deliver reasonable advantages, but have to be embedded into a larger concept. Also, the occurring problems like higher costs, should not be underestimated. Another problem is, that only tests working at a data centre can be performed there. This leads to the necessity of standardised and parameter driven quality checks, which were already motivated in section 2.6.6.

5.3.4 Peer review of data

For traditional publications, the peer review system has emerged as a basic requirement for accepting publications within the scientific community. When new forms of scientific publications, like the primary data publication are developed, it is an arising question, if there is the possibility of introducing such a system. This problem will be discussed in this section.

In section 5.3.4.1, the necessities and problems in the development of a peer review of primary data publications will be discussed. This is followed in section 5.3.4.2 by some comments on whether the presented quality evaluation concept in section 2.6.7 is a possible solution for this task. The necessary developments and standardisations to make it an acceptable option will be outlined. The last section 5.3.4.3 will take a look at alternatives and discuss whether a peer review system of data is needed at all.

5.3.4.1 Necessities and problems

This section explains, why a peer review of data concept is essential for data publication and which problems arise in the development of such a concept.

Doubts on the need of a peer review of data, base on the fact, that peer review is not unquestionable in modern science. New forms of technology, like they were shown with the pre-publications in the quality assurance process (section 2.2.3) start to threat the position of peer review in some sciences (Campanario

[1998b]). Also, the reason for the introduction of peer review in text publications might be not given in the current situation. The introduction was mainly driven by missing knowledge of a scientific field of the responsible editor (Burnham [1990]). Since the infrastructure of data publication is currently building up and the scientists working in data centres like the World Data Centres are experts in their field, there might be no actual need for such a system. Like it was described in section 2.2.1, not every journal used peer review for a long time and their publications were still accepted as scientific. Therefore, the question arises, why the development of a peer review system for data is necessary at all, although it might not be needed any more in modern sciences.

The necessity for peer review of data is mainly rooted in the fact, that such procedures are well established in traditional publications (see also section 2.2.1). The DFG does not see any possible alternative to a well performed peer review process for scientific publications (Deutsche Forschungsgemeinschaft [1998], p. 11). Cassella and Calvi [2010] cite a survey of Mark Ware, which claims that 93 % of scholars think that peer review is necessary. One reason is given by Spier [2002], who said: "In a world where knowledge is being made available at a rate of millions of pages per day, it is comforting to know that some subset of that knowledge or science has been critically examined so that, were we to use it in our thinking of for our work, we would be less likely to have wasted our time."

When it comes to data publication, there are some calls for the establishment of a peer review of such publications. The American Geophysical Union [1997] is for example "[...] encouraging peer-review of such publication". Officially, data is part of the traditional peer review at the AGU journals, if datasets are the basis of the research. They ask their reviewers to comment on their data (AGU Publications Committee [1993]), but the exact form of the procedure is not specified.

The status of non-peer-reviewed datasets is also not defined yet. Parsons et al. [2010] for example cited a suggestion that those data should be treated like a conference presentation. That the status of datasets are not yet defined is problematic, especially when the data is used to draw conclusions from it. This was also remarked by the InterAcademy Council, which reviewed the assessment report of the Intergovernmental Panel on Climate Change (IPCC). They call for guidelines on handling non-peer-reviewed literature, which include "[...] observational data sets, and model output" (InterAcademy Council 2010, p. 63).

Other reasons can be found in the fact that even if data is reviewed within traditional publications, the review process is not necessarily well done. An argument for this can be found in the time, that is currently used by a reviewer to review a paper. Lock and Smith [1990] collected some studies and determined numbers between 2 and 4 hours for medicine and around 6 to 8 hours for mathematics and physics. Yankauer [1990] estimated similar numbers. Therefore, the time for reviewing a paper is short, especially, when an additional review on datasets is required, like the AGU does. Similar can be expected for data journals. Just like the traditional publications it is a free form text to be evaluated and reviewed. If a shorter analysis in the data paper leads to a longer evaluation of the datasets itself, it still has to be evaluated systematically. The ESSD for example asks for the application of statistical tests to the data, but only if possible (ESSD [2012a]). This might lead to the application of tests only to subsets of the data, which is of course better than nothing, but not enough. Only if the whole dataset is checked, it can be seen as peer reviewed. Just like for text publications it is expected, that everything is checked properly. Therefore, it is necessary to develop a peer review system of data. Without such a system the high standard of scientific work, that was established in publications in the past, might be threatened.

A lot of scientists see peer review as a quality measure (Campanario [1998a]). If a dataset is declared as peer reviewed, some might assume it free of errors or at least that most problems within the datasets are documented. This leads to the conclusion, that as much time as possible of a reviewer should be used to search for undocumented problems within the dataset. The view as a quality measure can also be used as an argument for that the primary data publication has a right to exist besides the traditional publication and the data journal. That both, standalone data publication and overlay journals, like the data journals, are a good possibility to scientifically publish data was also motivated by Lawrence et al. [2011].

It has also to be remembered, that new people get attracted by new availability of data. Data publication, like primary data publication and data journals, might make data accessible, what were not simply available for a broad public before. This might attract people, who are no experts of the field, to perform their own work with it (Overpeck et al. [2011]). Therefore, additional care is necessary. When a quality statement like a peer review is assigned to such datasets, it really has to be checked properly.

Problems can be seen in the complexity of a data peer review. It is complicate to check large amounts of data and treat every part of the dataset equally. The process has to emerge from data publishers and the scientific community, since it has to gain recognition from all parties within the scientific process (Parsons et al. [2010]). In the next section, it will be investigated, whether it might be possible to generate a peer review on data with the help of the quality evaluation system.

5.3.4.2 Is quality evaluation a possible solution?

Like it was shown in the last section, a peer review of data has to be simple and effective. Under the assumption that a reviewer would need between 6 to 10 hours for a review of a publication, the question arises, which techniques have to be developed for such a system. A possible solution may be quality evaluation. It was introduced in section 2.6.7, used in an application in section 4.4 and discussed in section 5.2. The basic task of the system is to automatise the evaluation of quality checks under the guidance of the knowledge of experts. As it was discussed in section 5.2.1, the quality evaluation system has the ability to generate this connection effectively, but at the same time it has a lot of disadvantages. Subjectivity is one of them, although it is no critical one for peer review. At this point it is requested that a reviewer uses his/her own personal view based on his/her knowledge to make statements on the quality of the publication. Still, the question, how the quality of a dataset can be defined, could lead to larger problems. As long as there does not exist any standard for what a quality of a given percentage means, it is practically useless to utilise such an estimate as an absolute quality estimation. Another problem is that the flexibility of such a system provides the opportunity of manipulating the results. This could happen by the choice of improper parameters of tests together with questionable priors. Therefore it is questionable, if it is beneficial to publish a result of a quality evaluation alongside the published data. The only way to minimise these problems is to standardise everything: tests, parameters, priors and weightings. This standardisation can be done either by organisations or by the evaluation of similar datasets with given tests and parameters and estimate the priors by the knowledge, how well the qualities of these datasets are. The latter would lead to the problem of how to define similar datasets. The automatisation to categorise the datasets and allow a automatic connection to the reasonable sets of tests, parameters, priors and weightings, is a complicated but necessary task. The aim is a really effective and independent automatised review system of data in the future. If this is given and several similar datasets are evaluated with the system above, the information of the quality evaluation may be usable. Quality evaluation itself is not really a reasonable tool for the generation of absolute results. Still, relative results might help within the peer review process.

The principle can be seen in the application in section 4.4. Here, the datasets of different days of more than one year are compared to each other. With this, it might be possible to quickly find out problematic datasets. Of course, under the condition that the environment for this evaluation, like parameters and priors is chosen reasonably. Resulting from that, quality evaluation might be used as a tool for the reviewer to make the decision, with which subset of data he/she should start his/her investigation within a peer review system. It also helps the reviewer to get a relatively quick overview over the results of the

performed checks, without checking for example several hundred plots. Since the documentation of such a system is relatively simple, it fulfils the criteria of transparency, if the results of the quality estimation are published (see also section 5.3.2). The publication should not only list the resulting percentages, but also explain the whole process including the results of the tests and the chosen priors.
A further requirement for an acceptable peer review in science is the definition of the whole process of the peer review. Especially the selection of the reviewers is important. This could be organised similarly to the traditional publication and performed under the guidance of special journals or data centres (Lawrence et al. [2011]). The latter is of great importance, since the selection of independent reviewers and editors is a fundamental requirement for peer review of a publication in general (Campanario [1998a], Parsons et al. [2010]).
The existing alternatives to the peer review of data are briefly discussed in the next section.

5.3.4.3 Other solutions for these problems

Currently, a couple of ways to perform a peer review on data are used or discussed in the literature. A working system is for example the peer review process of the NASA Planetary Data System (NASA PDS [2012]). It defines a review system for focussing on the technical quality, like completeness and documentation, of data. In the guidelines it is defined, how the reviewers have to be its selected and what their tasks are. Since the reviewers originate from NASA itself, this is no classical peer review. Also the data itself, from the view of its content, is in accordance with these guidelines not necessarily under investigation. Therefore, it is comparable to the review guidelines of the ESSD, which were mentioned in section 5.3.4.1.
The other system, that is widely discussed as an alternative to real peer review systems, is the trust in the use of a dataset as an indicator of its quality. The assumption here is, that a dataset has a good quality, when it is cited by other scientists. That this is problematic, was already discussed by Parsons et al. [2010], since it might be possible that the dataset is cited as a bad example. Then, the assumption would not hold any longer and the system would become compromised.
A last possible solution would be the annotation of datasets, as it is done today in social networks (Quadt et al. [2012]). It has similarities to an open review system (Benos et al. [2007]) and might be a practical solution until a real peer review system is established.
Up to this day, there is no such solution for the problem of peer review available, even if it is necessary to guarantee an effective science in the future. The approach of the quality evaluation is only a step towards this goal and needs further tests to estimate its effectiveness. It is heavily dependent on standardisation of several crucial factors, which are not in sight up to now, just like an effective peer review process of data itself.

6 Conclusion

This thesis has shown, how a quality assurance of primary data publication may be structured and which procedures and methods have to be defined in order to make it comparable to existent scientific publication processes. The latter was illustrated by a description of the role of publications in general in science and a brief overview of the history. With this as a background, a scientific working scheme has been developed, which also covers the primary data publication. A further look was taken on the role and kinds of quality assurances in the different forms of publications. In that context, three different quality assurance processes, which are designed as a quality assurance on methods, metadata and data, have been developed. After a brief overview on the two others, the quality assurance on primary data was investigated in more detail. Therefore, a test-based concept, including guidelines for documentation, was developed. Part of this concept are general quality checks, which only depend on a defined number of parameters. Additionally, a proposal for an automatised analysis of quality checks, the quality evaluation procedure, was shown. It allows, with the help of priors given by experts, to interpret the results of the quality checks for even a large number of datasets.

To show the possibilities of general quality checks, four different types were introduced afterwards. The first two, the tests by Meek and Hatfield and distribution based tests, were already used before in quality control literature and are therefore only briefly presented. Some minor and major enhancements, shown in this thesis, allow them to have an even higher flexibility and make them usable in more applications. The histogram test is a new test, which allows with a new approach, to access the quality of datasets in a very flexible way. It has been shown, that with different modifications, implemented by using different kinds of measures for the difference between one dimensional histograms, it becomes possible to detect different kinds of inconsistencies within datasets. The last introduced general test is the change point detection method developed by Dose and Menzel. It allows to check different kinds of models, which are regressed to the data, and to evaluate their success afterwards. The success is measured by means of probability and by allowing to use a large number of different models, what leads to a high flexibility. Different types of models were used to process an inter-comparison test with eight different types of change point detection methods, that are used in homogenisation of climate time series. In addition, the potential to combine the change point method with other quality checks was shown.

Afterwards, these quality checks were used in some applications of meteorological and climatological datasets. Therefore, the detection of rounding within data was shown by an example of a wind measurement. Also, the analysis of model data of the NCEP and ERA40 reanalyses was performed and a comparison of the base and gridded data of the CRU temperature reconstruction was done. In a last application, data of a climate station, which measures 15 different meteorological variables, were analysed with the help of the quality evaluation procedure.

The following discussion of the methods and procedures was divided into three different parts. The first part took a further look on the general quality checks and showed their risks and potentials. Both checks, the histogram test and the change point detection method by Dose and Menzel, have been proven very successful for their application in quality assurance processes. Afterwards, the quality evaluation procedure was further investigated. It was determined, that it might be a useful tool in quality assurance processes, though it depends on a high grade of standardisation of the used tests, parameters, priors and weightings. The third part discussed the data publication as a whole and the possible introduction of the here presented methods into a data centre. It has been shown, that this new type of publication has a high potential for the future, but depends on further developments to make it effective, scalable and comparable to the actual accepted publications in science. At last, the possibility of a peer review of primary data publication was considered. It was determined, that despite the need of an effective concept for such a procedure, the problems and risks are currently not solved yet. The quality evaluation concept developed in this thesis can contribute as an important tool within a peer review system. Still, due to its limitations and the non-available standardisation of methods, it cannot represent such a system as a whole. For the above explained reasons, further research is required in this field of science.

The here shown procedures allow the introduction of a primary data publication system, which can be seen in big parts as comparable to the established traditional forms of scientific publications. Some of those were already tested in practice in the DFG research project "Publikation Umweltdaten". Nevertheless, especially the integration of the quality checks into a data centre and the testing of the here presented approaches to peer review of data are complicated tasks, which require effort in the same amount of the already conducted project. Furthermore, the expansion of the here provided methods to other types of data and other scientific fields would allow interesting applications of the results of this thesis. Some smaller possible enhancements of the general quality checks were already mentioned in the discussion of the methods. The further investigation of the analysis of the results of the histogram test by means of complex networks or the application of the methods of this thesis in relative homogenisation, are some of them. In addition, looking for new general quality checks may be worthwhile. For example, to develop a quality check for outliers, which takes extreme value theory into account, would be useful for quality assurance of meteorological and climatological data.

It was shown, that scientists as a user, have to influence the development of new types of publications to fit them for their use in their daily work. Since currently a lot of new types emerge and established standards, like peer review, are threatened to get lost, procedures to guarantee the availability and the quality of publications have to be developed and applied. Data availability by publications is without doubt a crucial chance for scientists in many fields. Still, without the introduction of new forms and standards there is a high risk for that the potential of quality assurances will not be used completely.

A Change point model

In section 3.4, change point models were presented. In this part of the appendix, some additional information is given, which completes the information explained in the main part of this thesis. First, the results of the calibration process for the Dose and Menzel methods from section 3.4.3.4 for the different thresholds is shown in section A.1. Afterwards, some remarks on the SNHT method are made in section A.2 They are followed by an overview of the results of the inter-comparison test in section A.3.

A.1 Calibration of the method by Dose and Menzel

In section 3.4.3.4, the method of Dose and Menzel was calibrated for different regression models and different thresholds. In the figures 3.10 to 3.12 were the results of the 95% threshold shown. In this appendix, the results for the 50% and the 99% threshold are shown for the three different types of regression models. The calibrated γ for each model was already presented in table 3.2. In the following, the first figures are displayed for the 50% threshold. This are the flat model in figure A.1 and the normal and original model in figure A.2. For the 90% model the flat and the normal model are shown in figure A.3 and the original model in figure A.4.

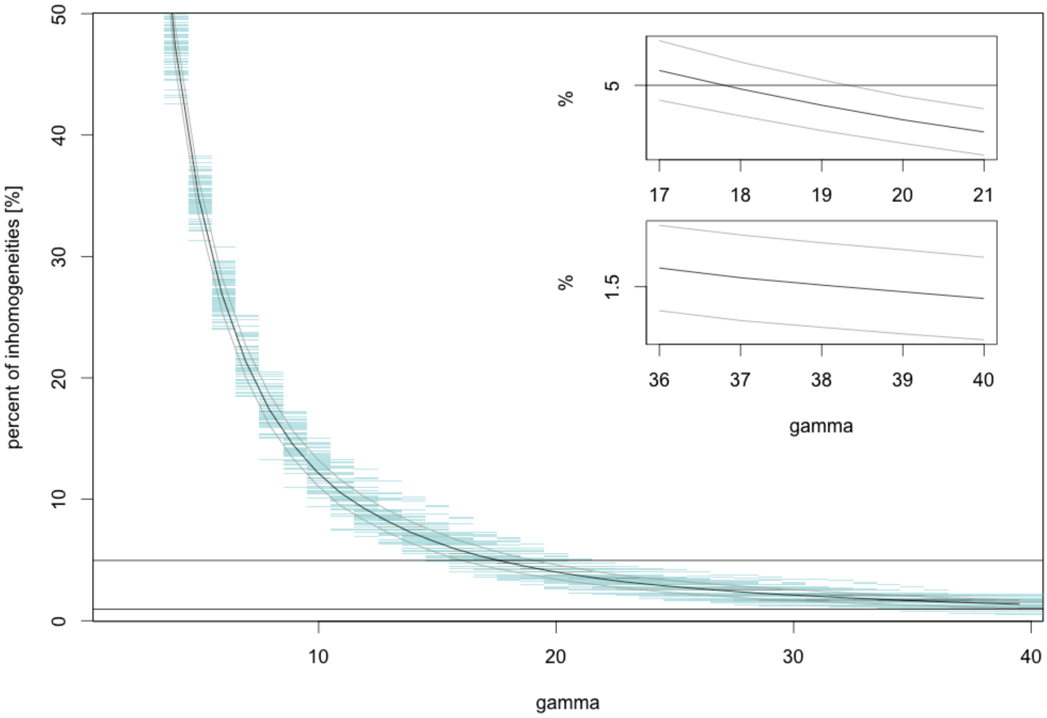
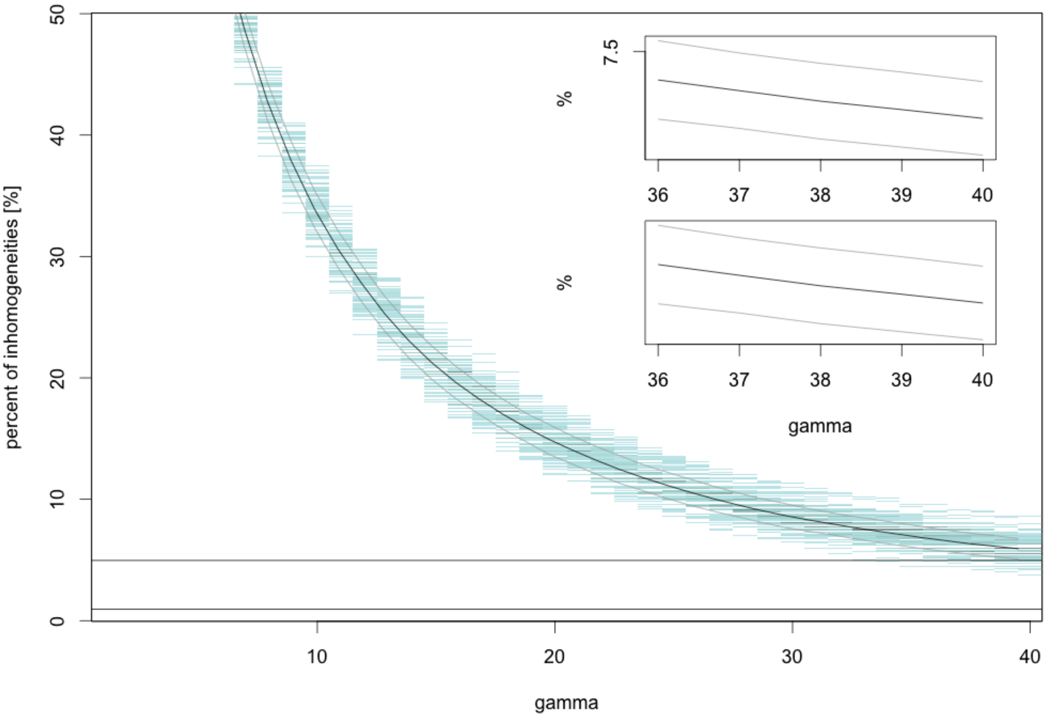


Figure A.1: Justification of γ for the flat model with a 50% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.



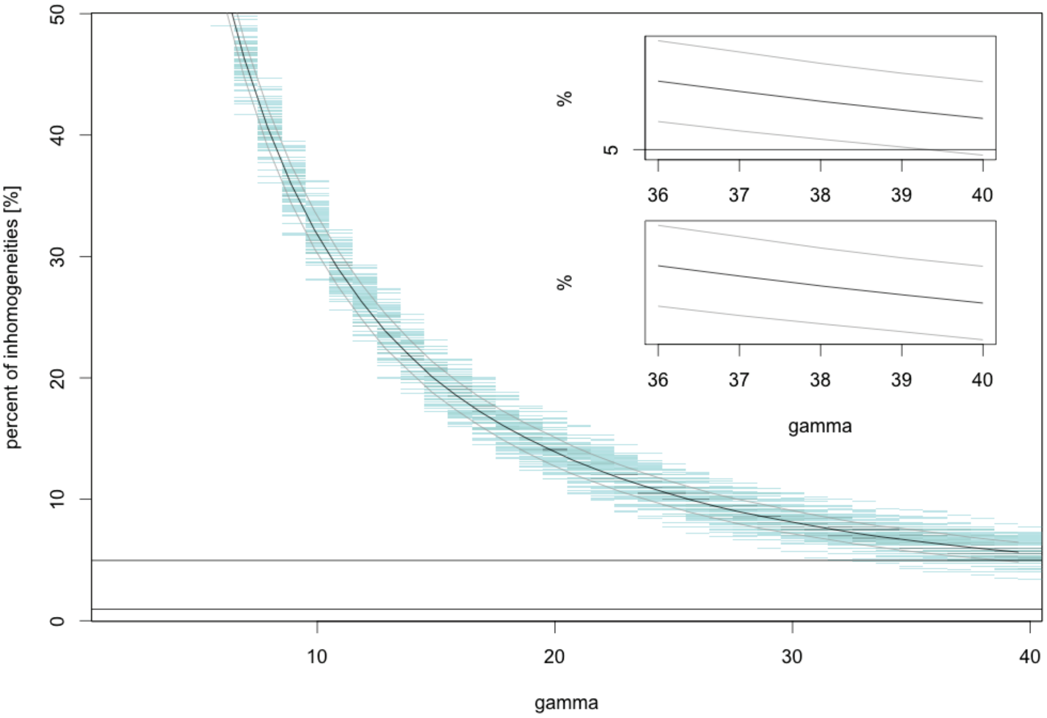
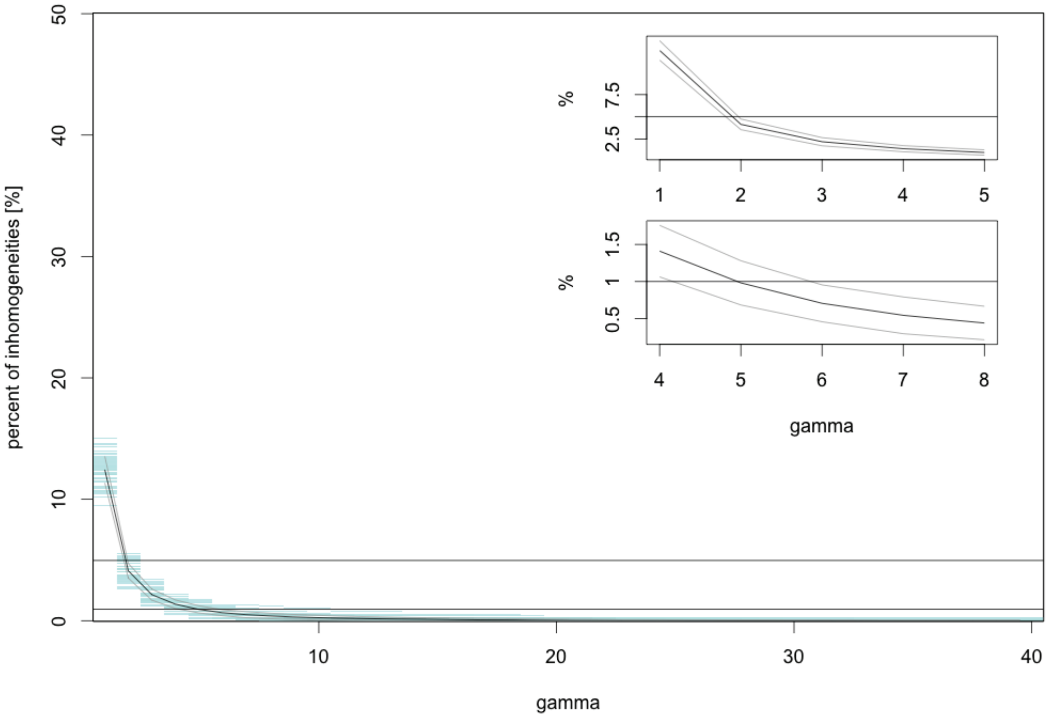


Figure A.2: Justification of γ for the normal (upper figure) and original (lower figure) model with a 50% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.



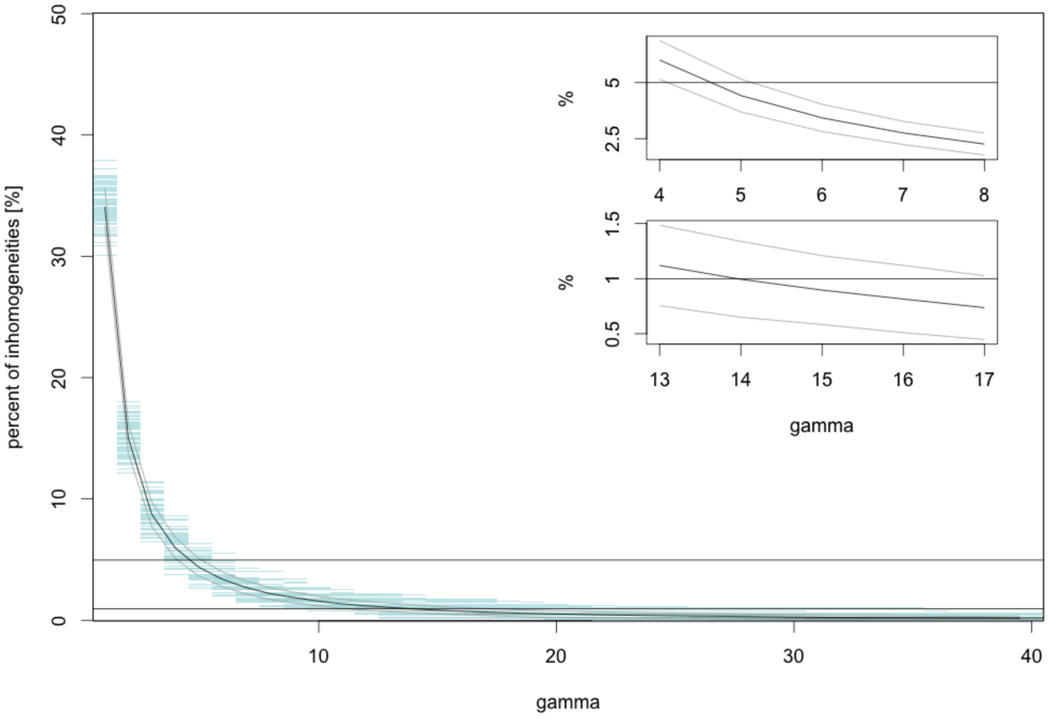


Figure A.3: Justification of γ for the flat (upper figure) and normal (lower figure) model with a 99% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.

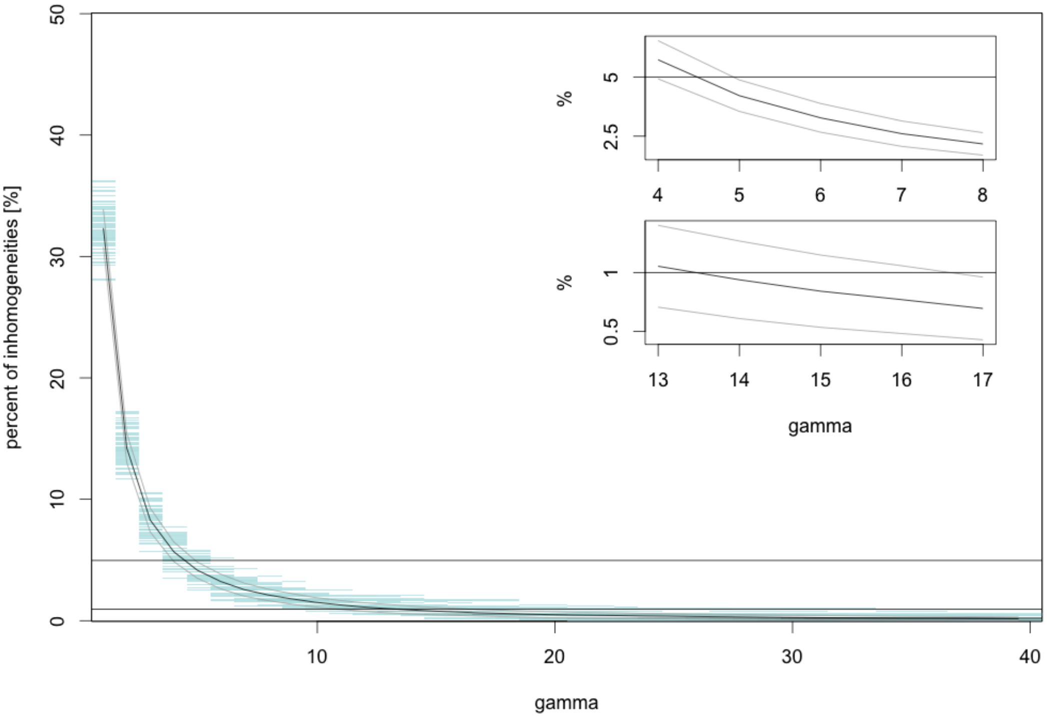


Figure A.4: Justification of γ for the original model with a 99% threshold. On the x-axis, the parameter γ is shown, on the y-axis, the percentage of inhomogeneities of 100000 homogeneous vectors. The black line indicates the mean, gray the standard deviations around the mean. The gray shadings behind the lines show the results for 100 packages of 1000 vectors. Subplots show the lower deviation of the 5% (top) and 1% (bottom) significance level.

A.2 SNHT

The SNHT method, which was presented in section 3.4.4.1 needs some additional remarks to fully explain the used procedures. It starts with some missing parameters for the equation, which define the SNHT method, are explained in section A.2.1. Afterwards the procedure to generate the critical values by interpolation for different length of vectors under consideration is shown in section A.2.2.

A.2.1 Parameters of the definition

The remaining parameters for the description of the SNHT method in section 3.4.4.1 for equation 3.55 are given as follows:

$$\mu_{SwT,1} = \frac{a_s \overline{z_{1:m_{chp}}} + S_{SwT,ze} - S_{SwT,zl} \cdot S_{SwT,se}}{a_s + S_{SwT,se} + S_{SwT,zk} \cdot S_{SwT,se}} \tag{A.1}$$

$$\mu_{SwT,2} = \mu_{SwT,1} S_{SwT,zk} + S_{SwT,zl} \tag{A.2}$$

$$S_{SwT,s} = \sum_{i=a_s+1}^{a_e} \frac{(i-a_s)^2}{(a_e-a_s)^2} \tag{A.3}$$

$$S_{SwT,e} = \sum_{i=a_s+1}^{a_e} \frac{(a_e-i)^2}{(a_e-a_s)^2} \tag{A.4}$$

$$S_{SwT,se} = \sum_{i=a_s+1}^{a_e} \frac{(i-a_s)(a_e-i)}{(a_e-a_s)^2} \tag{A.5}$$

$$S_{SwT,zs} = \sum_{i=a_s+1}^{a_e} \frac{z_i(i-a_s)}{(a_e-a_s)} \tag{A.6}$$

$$S_{SwT,ze} = \sum_{i=a_s+1}^{a_e} \frac{z_i(a_e-i)}{(a_e-a_s)} \tag{A.7}$$

$$S_{SwT,zl} = \frac{-S_{SwT,se}}{S_{SwT,s}+N-a_e} \tag{A.8}$$

$$S_{SwT,zk} = \frac{(N-a_e)\overline{z_{(m_{chp}+1):N}}+S_{SwT,zs}}{S_{SwT,s}+N-a_e} \tag{A.9}$$

A.2.2 Interpolation of the critical values

The critical values for the SNHT methods are given by Alexandersson and Moberg [1997] and Khaliq and Ouarda [2007]. In this thesis a transformation for interpolating the critical values for different length of the vectors N is used. Since a logarithm of the number of bins fits the values well, it is used to regress the transformation function to the given critical values. The results are shown in figure A.5. The red line indicates the interpolated values for the critical level of 95%, the original values by Alexandersson and Moberg [1997] are marked with crosses. Khaliq and Ouarda [2007] showed critical values for the critical level of 99%, which are marked with triangles and interpolated with a blue line.

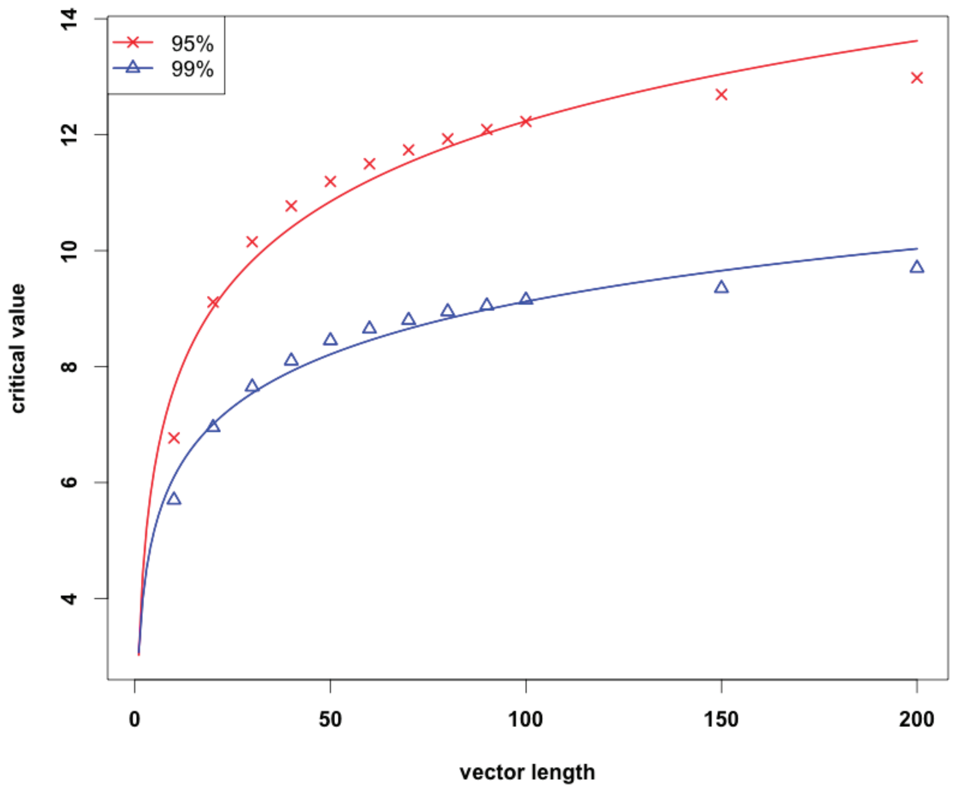


Figure A.5: Interpolation of the critical values of the SNHT method.

A.3 Results of the inter-comparison tests

In the following three tables, the results of the inter-comparison tests, presented in the sections 3.4.4 and 3.5, are shown. Table A.1 shows the results for the autoregressive process, table A.2 for the normal distributed and table A.3 for the gamma distributed test vectors. In each table, the elements of the test on homogeneous datasets, the results of the contingency table (see also table 3.5) of the test on inhomogeneous datasets and the calculated scores are shown.

Table A.1: Results of the autoregressive homogeneous and inhomogeneous test vectors for the intercomparison test.

method	hom.			change point detection					
name	inhom. [%]	a	b	c	d	log OR	HSS	entropy	CRPS
TPR	34.5	13668	3592	770	6970	3.539	0.629	1.547	0.147
MLR	16.5	11722	1877	2716	8685	2.994	0.627	1.671	0.140
SNHT/wo	2.7	13853	844	585	9718	5.608	0.882	1.294	0.053
SNHT/w	3.1	13743	1036	695	9526	5.203	0.857	1.339	0.072
ST	2.0	3044	185	11394	10377	2.707	0.169	1.465	0.283
WRS	50.9	13355	5289	1083	5273	2.509	0.448	1.627	0.167
Bayes/ref	31.8	14175	3483	263	7079	4.696	0.982	1.445	0.097
Bayes/noref	7.2	10555	730	3883	9832	3.600	0.636	1.621	0.143
DM/ref	2.9	13347	235	1091	10327	6.287	0.892	1.271	0.051
DM/r/n	5.9	13601	566	837	9996	5.659	0.885	1.294	0.063
DM/r/o	5.8	13625	554	813	10008	5.713	0.888	1.288	0.061
DM/noref	3.2	10968	410	3470	10152	4.360	0.694	1.542	0.123
DM/nr/n	3.2	11098	508	3340	10054	4.186	0.696	1.551	0.129
DM/nr/o	2.8	11143	487	3295	10075	4.248	0.701	1.544	0.126
DM99/ref	3.7	13404	282	1034	10280	6.159	0.893	1.272	0.051
DM99/r/n	5.9	13583	547	855	10015	5.673	0.885	1.294	0.063
DM99/r/o	5.4	13604	527	834	10035	5.739	0.889	1.287	0.061
DM99/noref	3.5	11120	453	3318	10109	4.315	0.702	1.540	0.120
DM99/nr/n	3.2	11047	479	3391	10083	4.228	0.694	1.549	0.129
DM99/nr/o	2.7	11092	457	3346	10105	4.295	0.699	1.542	0.127
DM50/ref	4.3	13477	326	961	10236	6.088	0.895	1.270	0.051
DM50/r/n	6.7	13693	688	745	9874	5.575	0.883	1.300	0.064
DM50/r/o	6.8	13720	677	718	9885	5.631	0.886	1.292	0.062
DM50/noref	4.2	11278	526	3160	10036	4.221	0.708	1.541	0.118
DM50/nr/n	3.9	11381	647	3057	9915	4.044	0.705	1.553	0.126
DM50/nr/o	4.3	11432	614	3006	9948	4.121	0.712	1.544	0.123
DM/ref/KLD	0.0	1403	127	13035	10435	2.180	0.073	1.288	0.318
DM/r/n/KLD	23.1	6147	2427	8291	8135	0.910	0.182	1.879	0.300
DM/r/o/KLD	8.0	2169	769	12269	9793	0.812	0.068	1.494	0.322
DM/noref/KLD	0.0	556	113	13882	10449	1.301	0.024	1.155	0.329
DM/nr/n/KLD	21.4	4498	2356	9940	8206	0.455	0.081	1.823	0.320
DM/nr/o/KLD	6.5	1300	694	13138	9868	0.341	0.021	1.382	0.331
DM/ref/JSD	0.3	660	29	13778	10533	2.856	0.037	1.149	0.327
DM/r/n/JSD	7.0	2912	643	11526	9919	1.360	0.124	1.541	0.311
DM/r/o/JSD	1.9	812	171	13626	10391	1.287	0.034	1.213	0.328
DM/noref/JSD	0.3	216	27	14222	10535	1.779	0.011	1.058	0.332
DM/nr/n/JSD	5.8	1591	586	12847	9976	0.749	0.047	1.402	0.326
DM/nr/o/JSD	1.1	383	149	14055	10413	0.644	0.011	1.130	0.333
DM/ref/RMS	0.0	125	6	14313	10556	2.732	0.007	1.027	0.333
DM/r/n/RMS	4.3	1415	371	13023	10191	1.093	0.054	1.343	0.324
DM/r/o/RMS	0.1	72	9	14366	10553	1.771	0.003	1.013	0.333
DM/noref/RMS	0.1	43	10	14395	10552	1.150	0.002	1.004	0.333
DM/nr/n/RMS	2.9	873	374	13565	10188	0.561	0.215	1.266	0.330
DM/nr/o/RMS	0.2	32	11	14406	10551	0.756	0.001	1.001	0.334
DM/ref/MS	0.0	126	5	14312	10557	2.923	0.009	1.027	0.333
DM/r/n/MS	4.7	1471	410	12967	10152	1.033	0.054	1.357	0.323
DM/r/o/MS	5.3	1315	534	13123	10028	0.632	0.035	1.359	0.327
DM/noref/MS	0.1	42	8	14396	10554	1.348	0.002	1.003	0.334
DM/nr/n/MS	3.4	923	424	13515	10138	0.490	0.020	1.283	0.330
DM/nr/o/MS	4.4	916	488	13522	10074	0.335	0.015	1.294	0.331
DM/ref/EMD	0.1	6874	10	7564	10552	6.866	0.434	1.564	0.194
DM/r/n/EMD	3.0	10984	392	3454	10170	4.413	0.697	1.538	0.126
DM/r/o/EMD	10.0	11692	1239	2746	9232	3.467	0.679	1.608	0.133
DM/noref/EMD	0.0	3508	10	10930	10552	5.825	0.213	1.449	0.268
DM/nr/n/EMD	4.9	7042	404	7396	10158	3.176	0.413	1.659	0.218
DM/nr/o/EMD	12.8	8312	1305	6126	9257	2.264	0.426	1.778	0.216

Table A.2: Results of the normal distributed homogeneous and inhomogeneous test vectors for the intercomparison test.

method	hom. test	change point detection							
name	inhom. [%]	a	b	c	d	log OR	HSS	entropy	CRPS
TPR	17.4	14438	1880	0	8682	∞	0.842	-	0.047
MLR	5.1	14435	562	3	10000	11.357	0.953	1.111	0.027
SNHT/wo	0.9	14420	72	18	10490	11.668	0.993	1.015	0.007
SNHT/w	0.4	14407	37	31	10525	11.792	0.994	1.010	0.007
ST	1.6	14425	145	13	10417	11.286	0.987	1.033	0.008
WRS	47.0	13728	4740	710	5822	3.168	0.529	1.565	0.136
Bayes/ref	98.0	14438	10342	0	220	∞	0.024	-	0.203
Bayes/noref	4.4	10596	400	3842	10162	4.249	0.667	1.563	0.133
DM/ref	0.2	14414	10	24	10552	13.359	0.997	0.997	0.012
DM/r/n	0.2	14438	20	0	10542	∞	0.998	-	0.020
DM/r/o	0.2	14438	20	0	10542	∞	0.998	-	0.019
DM/noref	1.7	10878	170	3560	10392	5.230	0.707	1.498	0.118
DM/nr/n	1.9	10821	208	3617	10354	5.003	0.699	1.511	0.127
DM/nr/o	1.8	10859	189	3579	10373	5.115	0.704	1.504	0.125
DM99/ref	0.2	14418	12	20	10550	13.359	0.997	0.997	0.012
DM99/r/n	0.1	14438	19	0	10543	∞	0.998	-	0.020
DM99/r/o	0.1	14438	19	0	10543	∞	0.998	-	0.020
DM99/noref	2.0	11038	197	3400	10365	5.141	0.717	1.494	0.115
DM99/nr/n	1.9	10757	198	3681	10364	5.030	0.696	1.512	0.128
DM99/nr/o	1.8	10808	183	3630	10379	5.129	0.701	1.506	0.126
DM50/ref	0.3	14419	14	19	10548	13.257	0.997	0.997	0.011
DM50/r/n	0.2	14438	23	0	10539	∞	0.998	-	0.019
DM50/r/o	0.2	14438	24	0	10538	∞	0.998	-	0.019
DM50/noref	2.2	11202	236	3236	10326	5.020	0.729	1.491	0.112
DM50/nr/n	2.4	11101	266	3337	10296	4.858	0.716	1.505	0.122
DM50/nr/o	2.1	11129	256	3309	10306	4.908	0.719	1.501	0.121
DM/ref/KLD	0.0	8287	137	6151	10425	4.630	0.521	1.593	0.191
DM/r/n/KLD	22.6	12776	2352	1662	8210	3.290	0.668	1.603	0.162
DM/r/o/KLD	7.5	10240	719	4198	9843	3.508	0.614	1.636	0.176
DM/noref/KLD	0.0	573	126	13865	10436	1.230	0.023	1.161	0.329
DM/nr/n/KLD	23.3	4398	2343	10040	8219	0.430	0.075	1.817	0.320
DM/nr/o/KLD	7.1	1287	692	13151	9870	0.333	0.020	1.380	0.331
DM/ref/JSD	0.1	6129	32	8309	10530	5.492	0.381	1.563	0.233
DM/r/n/JSD	5.5	10253	639	4185	9923	3.639	0.622	1.623	0.184
DM/r/o/JSD	2.3	7568	182	6870	10380	4.140	0.467	1.612	0.218
DM/noref/JSD	0.4	230	29	14208	10533	1.771	0.011	1.062	0.332
DM/nr/n/JSD	5.8	1554	610	12884	9952	0.677	0.431	1.402	0.327
DM/nr/o/JSD	1.0	376	154	14062	10408	0.592	0.010	1.130	0.332
DM/ref/RMS	0.2	2369	15	12069	10547	4.927	0.141	1.361	0.299
DM/r/n/RMS	3.9	6644	379	7794	10183	3.131	0.388	1.652	0.248
DM/r/o/RMS	0.1	2113	16	12325	10546	4.727	0.125	1.336	0.306
DM/noref/RMS	0.2	52	4	14386	10558	2.256	0.003	1.005	0.333
DM/nr/n/RMS	4.2	859	370	13579	10192	0.555	0.021	1.263	0.330
DM/nr/o/RMS	0.0	25	11	14413	10551	0.509	0.001	0.998	0.334
DM/ref/MS	0.2	2530	15	11908	10547	5.007	0.151	1.376	0.293
DM/r/n/MS	4.1	6803	408	7635	10154	3.099	0.396	1.658	0.243
DM/r/o/MS	4.1	6710	489	7728	10073	2.884	0.383	1.672	0.244
DM/noref/MS	0.2	52	4	14386	10558	2.256	0.003	1.005	0.333
DM/nr/n/MS	4.5	880	402	13558	10160	0.495	0.020	1.272	0.331
DM/nr/o/MS	4.7	861	468	13577	10094	0.313	0.013	1.281	0.331
DM/ref/EMD	0.3	11828	7	2610	10555	8.830	0.792	1.380	0.069
DM/r/n/EMD	4.0	14416	336	22	10226	9.901	0.971	1.078	0.023
DM/r/o/EMD	11.3	14336	1070	102	9492	7.128	0.903	1.217	0.040
DM/noref/EMD	0.2	3843	5	10494	10557	6.641	0.234	1.468	0.261
DM/nr/n/EMD	3.9	7312	352	7126	10210	3.393	0.435	1.649	0.210
DM/nr/o/EMD	12.9	8511	1145	5927	9417	2.469	0.454	1.756	0.207

Table A.3: Results of the gamma distributed homogeneous and inhomogeneous test vectors for the intercomparison test.

method	hom. test				change point detection				
name	inhom. [%]	a	b	c	d	log OR	HSS	entropy	CRPS
TPR	22.9	14438	2271	0	8291	∞	0.808	-	0.054
MLR	6.8	14438	601	0	9961	∞	0.950	-	0.027
SNHT/wo	1.8	14412	158	26	10404	10.505	0.985	1.041	0.008
SNHT/w	1.2	14397	81	41	10481	10.724	0.990	1.026	0.009
ST	1.0	14066	109	372	10453	8.196	0.961	1.117	0.018
WRS	45.7	13712	4848	726	5714	3.103	0.518	1.569	0.140
Bayes/ref	93.5	14438	9997	0	565	∞	0.061	-	0.203
Bayes/noref	0.1	4856	35	9582	10527	5.027	0.297	1.538	0.249
DM/ref	1.1	14403	67	35	10495	11.074	0.992	1.020	0.013
DM/r/n	2.7	14438	248	0	10314	∞	0.980	-	0.023
DM/r/o	2.6	14438	214	0	10348	∞	0.982	-	0.023
DM/noref	4.0	8560	387	5878	10175	3.645	0.520	1.641	0.181
DM/nr/n	7.1	8625	739	5813	9823	2.982	0.495	1.699	0.195
DM/nr/o	17.1	10307	1911	4131	8651	2.424	0.518	1.770	0.192
DM99/ref	1.2	14406	82	32	10480	10.960	0.991	1.023	0.013
DM99/r/n	2.6	14438	234	0	10328	∞	0.981	-	0.023
DM99/r/o	2.5	14438	203	0	10359	∞	0.983	-	0.023
DM99/noref	4.4	8763	442	5675	10120	3.565	0.530	1.647	0.178
DM99/nr/n	7.1	8543	723	5895	9839	2.982	0.491	1.698	0.196
DM99/nr/o	16.3	10227	1844	4211	8718	2.440	0.518	1.768	0.192
DM50/ref	1.3	14407	91	31	10471	10.887	0.990	1.026	0.013
DM50/r/n	3.2	14438	289	0	10273	∞	0.976	-	0.023
DM50/r/o	2.8	14438	251	0	10311	∞	0.979	-	0.022
DM50/noref	5.3	8970	503	5468	10059	3.491	0.540	1.652	0.175
DM50/nr/n	7.9	8968	872	5470	9690	2.902	0.509	1.709	0.192
DM50/nr/o	21.0	10719	2306	3719	8256	2.334	0.515	1.778	0.193
DM/ref/KLD	0.0	8008	122	6430	10440	4.669	0.503	1.594	0.197
DM/r/n/KLD	23.3	12593	2343	1845	8219	3.176	0.655	1.624	0.166
DM/r/o/KLD	10.3	9946	1007	4492	9555	3.045	0.568	1.691	0.188
DM/noref/KLD	0.0	594	139	13844	10423	1.169	0.024	1.168	0.329
DM/nr/n/KLD	24.8	4551	2408	9887	8154	0.444	0.080	1.829	0.320
DM/nr/o/KLD	10.1	1834	1030	12604	9532	0.298	0.026	1.495	0.330
DM/ref/JSD	0.7	5778	19	8660	10543	5.914	0.359	1.551	0.239
DM/r/n/JSD	6.4	9972	576	4466	9986	3.656	0.606	1.627	0.188
DM/r/o/JSD	2.4	7144	230	7294	10332	3.784	0.434	1.624	0.228
DM/noref/JSD	0.2	240	34	14198	10528	1.655	0.011	1.066	0.332
DM/nr/n/JSD	6.6	1615	611	12823	9951	0.718	0.047	1.409	0.326
DM/nr/o/JSD	2.8	543	225	13895	10337	0.585	0.014	1.179	0.332
DM/ref/RMS	0.2	2121	7	12317	10555	5.559	0.126	1.334	0.302
DM/r/n/RMS	5.2	6200	404	8238	10158	2.940	0.356	1.651	0.255
DM/r/o/RMS	0.3	1819	26	12619	10536	4.068	0.106	1.309	0.311
DM/noref/RMS	0.2	50	12	14388	10550	1.117	0.002	1.007	0.334
DM/nr/n/RMS	4.0	931	384	13507	10178	0.603	0.241	1.277	0.330
DM/nr/o/RMS	0.2	60	19	14378	10543	0.840	0.002	1.013	0.334
DM/ref/MS	0.2	2250	6	12188	10556	5.783	0.134	1.346	0.298
DM/r/n/MS	5.1	6397	427	8041	10135	2.938	0.367	1.658	0.248
DM/r/o/MS	7.3	6379	701	8059	9861	2.410	0.343	1.703	0.253
DM/noref/MS	0.1	53	11	14385	10551	1.262	0.002	1.008	0.334
DM/nr/n/MS	3.9	1012	418	13426	10144	0.604	0.026	1.296	0.329
DM/nr/o/MS	5.1	1157	629	13281	9933	0.320	0.018	1.353	0.331
DM/ref/EMD	0.2	11629	6	2809	10556	8.893	0.777	1.396	0.074
DM/r/n/EMD	3.4	14402	427	36	10135	9.159	0.962	1.100	0.026
DM/r/o/EMD	13.0	14308	1377	130	9185	6.599	0.874	1.261	0.049
DM/noref/EMD	0.1	2253	12	12185	10550	5.091	0.134	1.349	0.293
DM/nr/n/EMD	3.5	5267	431	9171	10131	2.603	0.292	1.633	0.255
DM/nr/o/EMD	12.9	6719	1382	7719	9180	1.755	0.310	1.795	0.251

B Quality evaluation

In section 4.4 the application of quality evaluation on the data of a climate station in Hohenheim was shown. In the following two tables B.1 and B.2, the settings for the tests, parameters, priors and weightings are given. Used are the LIM and ROC static tests and the NOC test. The description for the meteorological variables can be found in table 4.3. For the weightings and prior types the implementation was given in section 4.4.1.

Table B.1: Used tests, parameters, priors and weightings for the application of quality evaluation procedure on the climate station in Hohenheim (part 1).

Meteorol. variable	Test	Parameter 1	Parameter 2	Weighting	Prior type
TimeFromStart	ROC	downward-value: 0	upward-value:	3	A
T200	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -40.0	max-value: 40.0	2	B
		min-value: -25.0	max-value: 35.0	1	C
		min-value: -20.0	max-value: 30.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
RH200	LIM	min-value: 0	max-value: 100	3	A
		min-value: 20.0	max-value: 98	2	B
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
T005	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -40.0	max-value: 40.0	2	B
		min-value: -25.0	max-value: 35.0	1	C
		min-value: -20.0	max-value: 30.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
WD1000	LIM	min-value: 0	max-value: 360	3	A
	NOC	max-return-value: 20		3	A
		max-return-value: 15		2	B
		max-return-value: 10		1	C
		max-return-value: 5		1	D
WS1000	LIM	min-value: 0	max-value: 113.2	3	A
		min-value: 0	max-value: 80.0	3	A
		min-value: 0	max-value: 60.0	2	B
		min-value: 0	max-value: 40.0	1	C
		min-value: 0	max-value: 20.0	1	D
	NOC	max-return-value: 20		3	A
		max-return-value: 15		2	B
		max-return-value: 10		1	C
		max-return-value: 5		1	D
GR200	LIM	min-value: -100	max-value: 1367	3	A
		min-value: 0	max-value: 1200	2	D
RR200	LIM	min-value: -1367	max-value: 1367	3	A
		min-value: 0	max-value: 1200	2	D
NR200	LIM	min-value: -1367	max-value: 1367	3	A
		min-value: -100	max-value: 1200	2	D
RAIN	LIM	min-value: 0	max-value: 20	3	A
		min-value: 0	max-value: 15	2	C
		min-value: 0	max-value: 10	1	D

Table B.2: Used tests, parameters, priors and weightings for the application of quality evaluation procedure on the climate station in Hohenheim (part 2).

Meteorol. variable	Test	Parameter 1	Parameter 2	Weighting	Prior type
ST002	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -40.0	max-value: 40.0	2	B
		min-value: -25.0	max-value: 35.0	1	C
		min-value: -20.0	max-value: 30.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
ST005	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -40.0	max-value: 40.0	2	B
		min-value: -25.0	max-value: 35.0	1	C
		min-value: -20.0	max-value: 30.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
ST010	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -40.0	max-value: 40.0	2	B
		min-value: -25.0	max-value: 35.0	1	C
		min-value: -20.0	max-value: 30.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
ST020	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -30.0	max-value: 30.0	2	B
		min-value: -10.0	max-value: 25.0	1	C
		min-value: -5.0	max-value: 20.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D
ST050	LIM	min-value: -89.2	max-value: 57.8	3	A
		min-value: -30.0	max-value: 30.0	2	B
		min-value: -10.0	max-value: 25.0	1	C
		min-value: -5.0	max-value: 20.0	1	D
	ROC	downward-value: 10	upward-value: 10	3	A
		downward-value: 8	upward-value: 8	2	B
		downward-value: 6	upward-value: 6	1	C
		downward-value: 4	upward-value: 4	1	D

List of Variables

Variable	Short description	Section or Equation
a	hits	section 3.4.4.4
a_{max}	maximum threshold for the LIM/ROC static test	section 3.1
a_{min}	minimum threshold for the LIM/ROC static test	section 3.1
$a_{max,i}$	maximum threshold for the LIM/ROC dynamic test	section 3.1
$a_{min,i}$	minimum threshold for the LIM/ROC dynamic test	section 3.1
a_f	factor of prior for bin of histogram	section 3.3.2.1
a_i	observations in a bin of a histogram	section 3.3.2.1
a_k	intercept of a modified model	equation 3.41
a_p	prior for bin of histogram	equation 3.12
α	significance level	section 3.4.4.1.d)
$\mathbb{A}$	model matrix for functionals	section 3.4.2.1
$\mathbb{A}_C$	model matrix for a constant model	equation 3.36
$\mathbb{A}_{Cm}$	model matrix for a modified constant model	section 3.4.2.3
$\mathbb{A}_{CHP}$	model matrix for a original one change point model	section 3.4.2.2
$\mathbb{A}_{CHPf}$	model matrix for a flat one change point model	section 3.4.2.3
$\mathbb{A}_{CHPm}$	model matrix for a flat one change point model	section 3.4.2.3
$\mathbb{A}_L$	model matrix for a linear model	equation 3.38
$\mathbb{A}_{Lm}$	model matrix for a modified linear model	equation 3.44
B	beta function	equation 3.60
b	false alarm	section 3.4.4.4
b_k	slope of a modified model	equation 3.41
β	parameter for the standard deviation in Dose-Menzel	equation 3.4.2.1
c	miss	section 3.4.4.4
Γ	gamma function	section 3.4.2.1
γ	sensitivity parameter of Dose-Menzel	section 3.4.2.1
D_{DW}	parameter for the Durbin-Watson test	equation 3.50
D_{EM}	Earth Mover's Distance	equation 3.14
D_{JS}	Jenson-Shannon divergence	equation 3.13
D_{KL}	Kullback-Leibler divergence	equation 3.10
D_{MS}	Mean square measure	equation 3.15
D_{RMS}	Root mean square measure	equation 3.16
d	correct rejection	section 3.4.4.4
$\vec{d}$	data	section 3.4.2.1
δ	size of a step determined by sampling	section 3.4.4.1.f)
δ_{step}	size of a step	section 3.4.4.4
$\mathcal{E}$	exponential distribution	section 3.4.4.4
e_i	error of the MLR	equation 3.49

$\vec{\epsilon}$	error of a model fit to the data	section 3.4.2.1
F_X	cumulative distribution function	section 3.2.1
f_i	flag vector	section 3.1
f_X	probability density function	section 3.2.1
$\vec{f}$	functionals to fit to the data	section 3.4.2.1
h_i	bin of a histogram	equation 3.11
I	knowledge of an expert	section 2.6.7
I_B	general background	section 3.4.2.1
$\mathcal{J}$	set of model matrices	section 3.4.2.2
k	degree of freedom of a model	section 3.4.2.1
Λ	modification function	section 3.1.1
λ	singular values of the singular value decomposition	section 3.4.2.1
M_O	measurement operator	section 2.6.7
m_{block}	length of a block of a block window	section 3.2.2
m_{chp}	position of a change point	section 3.4.4
$\mu_{c,k}$	centralised kth moment	equation 3.8
μ_{diff}	mean of section with different basic distribution in histogram test	section 3.3.3.1
μ_k	k-th moment	equation 3.7
μ_{same}	mean of section with the same basic distribution in histogram test	section 3.3.3.1
$\mu_{SwT,i}$	parameter of the SNHT with reference series	section 3.4.4.1.c)
μ_x	mean of vector X	section 3.1.1
m_{slide}	length of a block of a sliding window	section 3.2.3
N	Number of elements of a vector	section 3.4.2.1
N_i	length of a section before or after a break point	section 3.4.4.1.d)
N_x	length of a section of a vector	section 3.4.4.1.e)
n_b	number of bins of a histogram	section 3.3.2.1
n_{chp}	number of change points	table 3.1
O	observation	section 2.6.7
p_i	element of a contingency table	table 3.4
Q	quality parameter in quality estimation	section 2.6.7
$\mathbb{Q}$	auxiliary matrix in Dose-Menzel	section 3.4.2.1
R	residue	section 3.4.2.1
$RSS_{TPR,i}$	residual sum of squares for the TPR method	section 3.4.4.1.a)
$RSS_{MLR,i}$	residual sum of squares for the MLR method	section 3.4.4.1.b)
r_{boot}	number of repetitions of a bootstrap	section 3.2.5
ρ	autocorrelation factor of	section 3.4.4.1.b)
S_{Bay}	parameter of the Bayes method	equation 3.58
S_{Ent}	entropy	equation 3.72
S_{HSS}	Heidke Skill Score	equation 3.71
S_{lOR}	log odds ratio	equation 3.70
$S_{SwT,i}$	parameter of the SNHT with reference series method	section 3.4.4.1.c)
$S_{WRS,i}$	parameter of the WRS method for defined sections of a vector	section 3.4.4.1.e)
$S_{WRS,x}$	parameter of the WRS method	section 3.4.4.1.e)
s	factor of standard deviations	section 3.1.1
s_b	size of block	section 3.3.2.1
σ_{DM}	standard deviation in the Dose-Menzel method	section 3.4.2.1

σ_{diff}	standard deviation of section with different basic distribution in histogram test	section 3.3.3.1
σ_i	standard devition of a deffined section of a vector	section 3.4.4.1.d)
σ_{same}	standard deviation of section with the same basic distribution in histogram test	section 3.3.3.1
σ_x	standard deviation of vector X	section 3.1.1
T	truth	section 2.6.7
t_{pos}	positioning of a steo determined by samling	section 3.4.4.4
θ	parameter of a test	section 2.6.7
U	unitary matrix of the singular value decomposition	section 3.4.2.1
$U_{MLR,m_{chp}}$	result value for the MLR method	equation 3.53
$U_{ST,m_{chp}}$	result value for the ST without reference method	equation 3.56
$U_{SwoT,m_{chp}}$	result value for the SNHT without reference method	equation 3.54
$U_{SwT,m_{chp}}$	result value for the SNHT with reference method	equation 3.55
$U_{TPR,m_{chp}}$	result value for the TPR method	equation 3.47
$U_{WRS,m_{chp}}$	result value for the WRS method	equation 3.56
V	orthogonal basis for the right-singular vectors of the singular value decomposition	section 3.4.2.1
V_s	multidimensional hypersphere	equation 3.24
X	dataset with elements x_i	section 3.1
X_i	time series	section 3.4.3
$\vec{x}$	time points of data point	section 3.4.2.1
$\overline{x_i}$	mean over a defined section of a time series	section 3.4.4.1.f)
x_i	element of a vector X	section 3.1
x_{orig}	original series	section 3.4.4.1.b)
x_p	p-th quantile	equation 3.9
x_{ref}	reference series	section 3.4.4.1.b)
x_{sd}	evaluation parameter for the histogram test	equation 3.18
Y_i	reference series	section 3.4.4.2
y_{sd}	factor of standard deviation	section 3.3.3.3
$\overline{z_i}$	mean over a defined section of a difference series	section 3.4.4.1.c)

List of Tables

List of Figures

Bibliography

AGU Council (2009). The importance of long-term preservation and accessibility of geophysical data. AGU Position Statement.

AGU Publications Committee (1993). Policy on referencing data in and archiving data for agu publications. Internet. Available from: `http://www.agu.org/pubs/authors/policies/data_policy.shtml` [cited 2012-03-06].

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., and Wieringa, J. (2003). Guidelines on climate metadata and homogenization. Technical Report WMO/TD No. 1186, World Meteorological Organization.

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of Climate*, 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*, 17:25–34.

American Geophysical Union (1997). Earth and space science data should be widely accessible in multiple formats and long-term preservation of data is an integral responsibility of scientists and sponsoring institutions. Internet. Available from: `http://www.agu.org/sci_pol/pdf/position_statements/AGU_Data_Statement.pdf` [cited 2012-06-03].

Bacon, F. (1620). *The New Organon*. Cambridge Texts in the History of Philosophy. Cambridge University Press, new (2000) edition.

Baker, N. L. (1992). Quality control for the navy operational atmospheric database. *Weather and Forecasting*, 7(2):250–261.

Beall, J. (2008). The weaknesses of full-text searching. *The Journal of Academic Librarianship*, 34:438–444.

Benos, D. J., Bashari, E., Chaves, J. M., Gaggar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., and Zotov, A. (2007). The ups and downs of peer review. *Advances in Physiology Education*, 31:145–152.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer.

Blakeslee, D. M. and Rumble Jr., J. (2003). The essentials of a database quality process. *Data Science Journal*, 2:35–46.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., Laurie, G., O'Neill, O., Rawlins, M., Janet, T., Vallance, P., and Walport, M. (2012). Science as an open enterprise. Technical report, The Royal Society.

Brase, J. (2009). Datacite - a global registration agency for research data. In *COINFO '09. Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009.*

Bretherton, F. P. and Singley, P. T. (1994). Metadata: a user's view. In *Seventh International Working Conference on Scientific and Statistical Database Management, 1994. Proceedings.*, pages 166–174.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research*, 111(D12106):1–21.

Bromwich, D. H. and Fogt, R. L. (2004). Strong trends in the skill of the era-40 and ncep–ncar reanalyses in the high and midlatitudes of the southern hemisphere, 1958–2001. *Journal of Climate*, 17:4603–4619.

Brown, H. (1972). History and the learned journal. *Journal of the History of Ideas*, 33(3):365–378.

Burnham, J. C. (1990). The evolution of editorial peer review. *Journal of the American Medical Association*, 263(10):1323–1329.

Campanario, J. M. (1998a). Peer review for journals as it stands today—part 1. *Science Communication*, 19(3):181–211.

Campanario, J. M. (1998b). Peer review for journals as it stands today—part 2. *Science Communication*, 19(4):277–306.

Canty, A. and Ripley, B. (2011). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-3.

Carlson, D. (2011). A lesson in sharing. *Nature*, 469:293.

Casadevall, A. and Fang, F. C. (2009). Is peer review censorship? *Infection and Immunity*, 77(4):1273–1274.

Casati, F., Marchese, M., Mirylenka, K., and Ragone, A. (2010). Reviewing peer review: a quantitative analysis of peer review. Technical report, Information Engineering and Computer Science.

Cassella, M. and Calvi, L. (2010). New journal models and publishing perspectives in the evolving digital environment. *IFLA Journal*, 36(1):7–15.

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.

Costello, M. J. (2009). Motivating online publication of data. *BioScience*, 59(5):418–427.

Crewell, S., Mech, M., Reinhardt, T., Selbach, C., Betz, H.-D., Brocard, E., Dick, G., O'Connor, E., Fischer, J., Hanisch, T., Hauf, T., Hünerbein, A., Delobbe, L., Mathes, A., Peters, G., Wernli, H., Wiegner, M., and Wulfmeyer, V. (2008). The general observation period 2007 within the priority program on quantitative precipitation forecasting: Concept and first results. *Meteorologische Zeitschrift*, 17(6):849–866.

DeGaetano, A. T. (2006). Attributes of several methods for detecting discontinuities in mean temperature series. *Journal of Climate*, 19:838–853.

del Barrio, E., Gine, E., and Matran, C. (1999). Central limit theorems for the wasserstein distance between the empirical and the true distributions. *The Annals of Propability*, 27(2):1009–1071.

Deutsche Forschungsgemeinschaft (1998). *Vorschläge zur Sicherung guter wissenschaftlicher Praxis*. Wiley-VCH.

Dose, V. (2009). Analysis of rare-event time series with application to caribbean hurricane data. *EPL (Europhysics Letters)*, 85(5):59001. Available from: `http://stacks.iop.org/0295-5075/85/i=5/a=59001`.

Dose, V. and Menzel, A. (2004). Bayesian analysis of climate change impacts in phenology. *Global Change Biology*, 10:259–272.

Dose, V. and Menzel, A. (2006). Bayesian correlation between temperature and blossom onset data. *Global Change Biology*, 12(8):1451–1459.

Drott, M. C. (2007). Open access. *Annual Review of Information Science and Technology*, 40:79–109.

Ducré-Robitaille, J.-F., Vincent, L. A., and Boulet, G. (2003). Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, 23:1087—1101.

Duerr, R. E., Downs, R. R., Tilmes, C., Barkstrom, B., Lenhardt, W. C., Glassy, J., Bermudez, L. E., and Slaughter, P. (2011). On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Information*, 4(3):139–160.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression i. *Biometrika*, 37(3/4):409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression ii. *Biometrika*, 38(1/2):159–177.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49(8):1615–1633.

Durre, I., Menne, M. J., and Vose, R. S. (2008). Strategies for evaluating quality assurance procedures. *Journal of Applied Meteorology and Climatology*, 47(6):1785–1791.

Düsterhus, A. (2011). qat: Quality assurance toolkit (version 0.5).

Düsterhus, A. and Hense, A. (2012). Advanced information criterion for environmental data quality assurance. *Advances in Science and Research*, 8:99–104.

Easterling, D. R. and Peterson, T. C. (1995). A new method for detecting undocumented discontinuities in climatological time series. *International Journal of Climatology*, 15:369–377.

Eaton, B., Gregory, J., Drach, B., Taylor, K. E., and Hankin, S. (2011). Netcdf climate and forecast (cf) metadata conventions. internet. Available from: `http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/cf-conventions-multi.html` [cited 2012-07-09].

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49:1858 – 1860.

ESSD (2012a). Essd - ms evaluation criteria. Internet. Available from: http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html [cited 2012-03-06].

ESSD (2012b). Review process & interactive public discussion. Internet. Available from: http://www.earth-system-science-data.net/review/review_process_and_interactive_public_discussion.html [cited 2012-03-06].

Fersht, A. (2009). The most influential journals: Impact factor and eigenfactor. *PNAS*, 106(17):6883–6884.

Fox, C., Levitin, A., and Redman, T. (1994). The notion of data and its quality dimensions. *Information Processing & Management*, 30:9–19.

Fox, J. and Weisberg, S. (2011). *car: An R Companion to Applied Regression.* Sage, Thousand Oaks CA, second edition. Available from: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Fuglede, B. and Topsøe, F. (2004). Jensen-shannon divergence and hilbert space embedding. *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, page 31.

Furrer, R., Nychka, D., and Sain, S. (2011). *fields: Tools for spatial data.* R package version 6.6.2. Available from: http://CRAN.R-project.org/package=fields.

Gandin, L. S. (1988). Complex quality control of meteorological observations. *Monthly Weather Review*, 116(5):1137–1156.

Giridharan, R., Lau, S., and Ganesan, S. (2005). Nocturnal heatislandeffect in urban residential developments of hong kong. *Energy and Buildings*, 37:964–971.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

GODAE (2007). Data definition. Internet. Available from: http://www.godae.org/Data-definition.html [cited 2012-05-02].

Gronell, A. and Wijffels, S. E. (2008). A semiautomated approach for quality controlling large historical ocean temperature archives. *JOURNAL OF ATMOSPHERIC AND OCEANIC TECHNOLOGY*, 25(6):990–1003.

Gullett, D. W., Vincent, L., and Sajecki, P. J. F. (1990). Testing homogeneity in temperature series at canadian climate stations. Technical Report CCC report 90-4, Climate Research Branch, Meteorological Service of Canada, Ontario, Canada.

Gura, T. (2002). Scientific publishing: Peer review, unmasked. *Nature*, 416:258–260.

Guralnick, R., Constable, H., Wieczorek, J., Moritz, C., and Peterson, A. T. (2009). Data's shameful neglect. *Nature*, 462(34):145.

Guttman, N. B. and Quayle, R. G. (1990). A review of cooperative temperature data validation. *Journal of Atmospheric and Oceanic Technology*, 7:334–339.

Hargens, L. L. (1990). Variation in journal peer review systems: Possible causes and consequences. *Journal of the American Medical Association*, 263(10):1348–1352.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72(357):180–186.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, 37:323–341.

Heidke, P. (1926). Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, 8:301–349.

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library trends*, 57(2):280–299.

Hemminger, B. M., Saelim, B., Sullivan, P. F., and Vision, T. J. (2007). Comparison of full-text searching to metadata searching for genes in two biomedical literature cohorts. *Journal of the American Society for Information Science and Technology*, 58:2341–2352.

Henneken, E. A. and Accomazzi, A. (2011). Linking to data - effect on citation rates in astronomy. *arXiv.org*, (arXiv:1111.3618).

Hense, A. and Quadt, F. (2011). Acquiring high quality research data. *D-Lib Magazine*, 17(1/2).

Hense, A. and Wulfmeyer, V. (2008). The german priority program spp1167 “quantitative precipitation forecast”. *Meteorologische Zeitschrift*, 17(6):703–705.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15:559–570.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, 56(3):495–504.

Højstrup, J. (1993). A statistical data screening procedure. *Measurement Science and Technology*, 4(2):153–157.

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., and Altman, D. G. (2010). Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *Trials*, 11(9).

Hubbard, K. G., Goddard, S., Sorensen, W. D., Wells, N., and Osugi, T. T. (2005). Performance of quality assurance procedures for an applied climate information system. *Journal of Atmospheric and Oceanic Technology*, 22:105–112.

Ingleby, N. B. and Lorenc, A. C. (1993). Bayesian quality-control using multivariate normal-distributions. *Quarterly Journal of the Royal Meteorological Society*, 119(513):1195–1225.

Jiménez, P. A., González-Rouco, J. F., Navarro, J., Montávez, J. P., and Garcia-Bustamante, E. (2010). Quality assurance of surface wind observations from automated weather stations. *Journal of Atmospheric and Oceanic Technology*, 27(7):1101–1122.

Jubb, M. (2012). Freedom of information in the uk and its implications for research in the higher education sector. *The International Journal of Digital Curation*, 7:57–71.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, B., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., and Dennis, J. (1996). The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77:437–472.

Karl, T. R. and Williams Jr., C. N. (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Applied Meteorology*, 26:1744–1763.

Khaliq, M. and Ouarda, T. B. M. J. (2007). On the critical values of the standard normal homogeneity test (snht). *International Journal of Climatology*, 27:681–687.

Killeen, T. (2012). Dear colleague letter - data citation. Internet. Available from: `http://www.nsf.gov/pubs/2012/nsf12058/nsf12058.jsp` [cited 2012-06-03].

Kinne, O. (1988). The scientific process - its links, functions and problems. *Naturwissenschaften*, 75:275–279.

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., and Fiorino, M. (2001). The ncep–ncar 50-year reanalysis: Monthly means cd-rom and documentation. *Bulletin of the American Meteorological Society*, 82(2):247–267.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M., Schindler, U., Sens, I., and Wächter, J. (2006). Data publication in the open access initiative. *Data Science Journal*, 5:79–83.

Knoll, E. (1990). The communities of scientists and journal peer review. *Journal of the American Medical Association*, 263(10):1330–1332.

Komsta, L. and Novomestky, F. (2011). *moments: Moments, cumulants, skewness, kurtosis and related tests*. R package version 0.12. Available from: `http://CRAN.R-project.org/package=moments`.

Kostoff, R. N. (2010). Expanded information retrieval using full-text searching. *Journal of Information Science*, 36(1):104–113.

Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. *Journal of the American Medical Association*, 263(10):1321–1322.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kunkel, K. E., Easterling, D. R., Hubbard, K., Redmond, K., Andsager, K., Kruk, M. C., and Spinar, M. L. (2005). Quality control of pre-1948 cooperative observer network data. *Journal of Atmospheric and Oceanic Technology*, 22:1691—1705.

Lancaster, F. W. (1995). The evolution of electronic publishing. *Library Trends*, 43(3):518–527.

Lang, D. T. (2011). *XML: Tools for parsing and generating XML within R and S-Plus.* R package version 3.4-3. Available from: `http://CRAN.R-project.org/package=XML`.

Lautenschlager, M., Toussaint, F., Thiemann, H., and Reinke, M. (1998). The cera-2 data model. Technical Report 15, Deutsches Klimarechenzentrum, Hamburg.

Lawrence, B., Jones, C., Matthews, B., Pepler, S., and Callaghan, S. (2011). Citation and peer review of data: Moving towards formal data publication. *The International Journal of Digital Curation*, 6:4–37.

Levina, E. and Bickel, P. (2001). The earth mover's distance is the mallows distance: some insights from statistics. *Eighth IEEE International Conference on Computer Vision, 2001. ICCV 2001. Proceedings.*, pages 251–256.

Lide, D. R. (2007). Data quality - more important than ever in the internet age. *Data Science Journal*, 6:154–155.

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145 – 151.

Lock, S. and Smith, J. (1990). What do peer reviewers do? *Journal of the American Medical Association*, 263(10):1341–1343.

Lorenc, A. C. and Hammon, O. (1988). Objective quality-control of observations using bayesian methods - theory, and a practical implementation. *Quarterly Journal of the Royal Meteorological Society*, 114(480):515–543.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141.

McNutt, R. A., Evans, A. T., Fletcher, R. H., and Fletcher, S. W. (1990). The effects of blinding on the quality of peer review: A randomized trial. *Journal of the American Medical Association*, 263(10):1371–1376.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., 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*, 88:1383–1394.

Meek, D. W. and Hatfield, J. L. (1994). Data quality checking for single station meteorological databases. *Agricultural and Forest Meteorology*, 69(1-2):85–109.

Menne, M. J. and Williams Jr., C. N. (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18:4271–4286.

Menzel, A. and Dose, V. (2005). Analysis of long-term time series of the beginning of flowering by bayesian function estimation. *Meteorologische Zeitschrift*, 14(3):429–436.

Met Office (2011). Land surface climate station records. Internet. Available from: `http://www.metoffice.gov.uk/research/climate/climate-monitoring/land-and-atmosphere/surface-station-records` [cited 2012-07-26].

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ecmwf ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122:73–119.

Monahan, A. H. (2006). The probability distribution of sea surface wind speeds. part ii: Dataset intercomparison and seasonal variability. *Journal of Climate*, 19:521–534.

Moreno, E., Casella, G., and Garcia-Ferrer, A. (2005). An objective bayesian analysis of the change point problem. *Stochastic Environmental Research and Risk Assessment*, 19(3):191–204.

NASA PDS (2012). The peer review process. Internet. Available from: `http://pds.nasa.gov/tools/peer-reviews.shtml` [cited 2012-07-15].

National Science Foundation (2011). Grant proposal guide. Internet. Available from: `http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_index.jsp` [cited 2012-08-07].

Oke, T. R. (1973). Citysize and the urban heat island. *Atmospheric Environment*, 7:769–779.

Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, 108:1–24.

Ouarda, T. B. M. J., Rasmussen, P. F., Cantin, J. F., Bobée, B., Laurence, R., and Hoang, V. D. (1999). Identification d'un réseau hydrométrique pour le suivi des modifications climatiques dans la province de québec. *Revue des sciences de l'eau / Journal of Water Science*, 12(2):425–448.

Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R. (2011). Climate data challenges in the 21st century. *Science*, 331(6018):700–702.

Page, E. S. (1955). A test for a change om parameter occuring at an unknown point. *Biometrika*, 42(3/4):523–527.

Palmer, T. N. (2012). You have free access to this contenttowards the probabilistic earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 138:841–1120.

Parsons, M. A., Duerr, R., and Minster, J.-B. (2010). Data citation and peer review. *Eos, Transactions American Geophysical Union*, 91(34):297–299.

Paskin, N. (2005). Digital object identifiers for scientific data. *Data Science Journal*, 4:12–20.

Pearl, J. (2000). *Causality: models, reasoning, and inference.* Cambridge University Press.

Perreault, L., Bernier, J., Bobée, B., and Parent, E. (2000). Bayesian change-point analysis in hydrometeorological time series. part 1. the normal model revisited. *Journal of Hydrology*, 235:221–241.

Perreault, L., Haché, M., Slivitzky, M., and Bobée, B. (1999). Detection of changes in precipitation and runoff over eastern canada and u.s. using a bayesian approach. *Stochastic Environmental Research and Risk Assessment*, 13(3):201–216.

Peterson, T. C. and Easterling, D. R. (1994). Creation of homogeneous composite climatological reference series. *International Journal of Climatology*, 14:671–679.

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

Pfeiffenberger, H. and Carlson, D. (2011). "earth system science data" (essd) — a peer reviewed journal for publication of data. *D-Lib Magazine*, 17(1/2).

Philosophical Transaction Staff (1665). The introduction. *Philosophical Transactions*, 1:1–2.

Pierce, D. (2011). ncdf: Interface to unidata netcdf data files: R package version 1.6.5. http://CRAN.R-project.org/package=ncdf.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Development Core Team (2011). *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-102.

Popper, K. R. (1934). *Logik der Forschung.* Mohr Siebeck, 11th (2005) edition.

Pöschl, U. (2010). Interactive open access publishing and public peer review: The effectiveness of transparency and self-regulation in scientific quality assurance. *IFLA Journal*, 36(1):40–46.

Quadt, F., Düsterhus, A., Höck, H., Lautenschlager, M., Hense, A. V., Hense, A. N., and Dames, M. (2012). Atarrabi – a workflow system for the publication of environmental data. in review.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available from: `http://www.R-project.org/`.

Rabin, J., Delon, J., and Gousseau, Y. (2008). Circular earth mover's distance for the comparison of local features. *19th International Conference on Pattern Recognition.*

Reek, T., Doty, S. R., and Owen, T. W. (1992). A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bulletin of the American Meteorological Society*, 73(6):753–762.

Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46:900–915.

Renear, A. H. and Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325:828.

Research Councils UK (2012). Research councils uk policy on access to research outputs. Internet. Available from: `http://www.rcuk.ac.uk/documents/documents/RCUK%20_Policy_on_Access_to_Research_Outputs.pdf` [cited 2012-08-19].

Rodionov, S. N. (2004). A sequential algorithm for testing climate regime shifts. *Geophys. Res. Lett.*, 31(9). Available from: `http://dx.doi.org/10.1029/2004GL019448`.

Rubner, Y., Puzicha, J., Tomasi, C., and Buhmann, J. M. (2001). Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, 84:25–43.

Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (ICCV'98)*, page 59.

Rubner, Y., Tomasi, C., and Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121.

Ruttenberg, S. and Rishbeth, H. (1994). World data centres — past, present and future. *Journal of Atmospheric and Terrestrial Physics*, 56(7):865–870.

Schleip, C. (2009). *Climate change detection in natural systems by Bayesian methods.* PhD thesis, TU München.

Schleip, C., Menzel, A., and Dose, V. (2009a). Bayesian analysis of changes in radiosonde atmospheric temperature. *International Journal of Climatology*, 29:629–641.

Schleip, C., Rutishauser, T., Luterbacher, J., and Menzel, A. (2008). Time series modeling and central european temperature impact assessment of phenological records over the last 250 years. *Journal of Geophysical Research*, 113:G04026.

Schleip, C., Sparks, T. H., Estrella, N., and Menzel, A. (2009b). Spatial variation in onset dates and trends in phenology across europe. *Climate Research*, 39:249–260.

Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., Rosenthal, N., and CASIMIR Rome Meeting participants (2009). Post-publication sharing of data and tools. *Nature*, 461:171–173.

Science Staff (2011). Challenges and opportunities. *Science*, 331:692–693.

Solow, A. R. (1987). Testing for climate change: An application of the two-phase regression model. *Journal of Climate and Applied Meteorology*, 26:1401–1405.

Spier, R. (2002). The history of the peer-review process. *Trends in Biotechnology*, 20:357–358.

Stephenson, D. B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, 15:221–232.

Strebel, D. E., Landis, D. R., Huemmrich, K. F., Newcomer, J. A., and Meeson, B. W. (1998). The fife data publication experiment. *Journal of the Atmospheric Sciences*, 55:1277–1283.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410:268–276.

Talbott, T. D., Schuchardt, K. L., Stephan, E. G., and Myers, J. D. (2006). Mapping physical formats to logical models to extract data and metadata: The defuddle parsing engine. *Provenance and annotation of data*, 4145:73–81.

Toronto International Data Release Workshop Authors (2009). Prepublication data sharing. *Nature*, 461:168–170.

Toussaint, F., Lautenschlager, M., and Luthardt, H. (2007). World data center for climate data—support for the ceop project in terms of model output. *Journal of the Meteorological Society of Japan*, 85A:475–485.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7(1):1–20. Available from: `http://dx.doi.org/10.1002/for.3980070102`.

Uppala, S. M., Kallberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Van De Berg, L., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Holm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., McNally, A. P., Mahfouf, J. F., Morcrette, J. J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The era-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

U.S. FGGE Project Office Staff (1978). *The Global Weather Experiment – Perspectives on Its Implementation and Exploitation.* The National Research Council.

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

Vickers, D. and Mahrt, L. (1997). Quality control and flux sampling problems for tower and aircraft data. *Journal of Atmospheric and Oceanic Technology*, 14(3):512–526.

Vigneron, V. (2006). Entropy-based principle and generalized contingency tables. In *ESANN'2006 proceedings - European Symposium on Artificial Neural Networks.* d-side publication.

Vincent, L. A. (1998). A technique for the identification of inhomogeneities in canadian temperature series. *Journal of Climate*, 11:1094–1104.

Von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Cliamte Research*. Cambridge University Press.

Wan, H., Wang, X. L. L., and Swail, V. R. (2007). A quality assurance system for canadian hourly pressure data. *Journal of Applied Meteorology and Climatology*, 46(11):1804–1817.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34.

Warner, S. (2005). The transformation of scholarly communication. *Learned Publishing*, 18(3):177–185.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2011a). *gplots: Various R programming tools for plotting data*. R package version 2.10.1. Available from: `http://CRAN.R-project.org/package=gplots`.

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., and Rogers, J. (2011b). *gdata: Various R programming tools for data manipulation*. R package version 2.8.2. Available from: `http://CRAN.R-project.org/package=gdata`.

Weibel, S. (1997). The dublin core: A simple content description model for electronic resources. *Bulletin of the American Society for Information Science and Technology*, 24:9–11.

Weller, A. C. (1990). Editorial peer review in us medical journals. *Journal of the American Medical Association*, 263(10):1344–1347.

Werman, M., Peleg, S., and Rosenfeld, A. (1985). A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32:328–336.

Williamson, J. (2005). *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford University Press.

WMO (2012). From data to products. Internet. Available from: `http://www.wmo.int/pages/prog/sat/dataandproducts_en.php` [cited 2012-05-02].

Wulfmeyer, V. and Henning-Müller, I. (2007). Climatological station hohenheim (since 1878). Internet. Available from: `http://gop.meteo.uni-koeln.de/gop/doku.php?id=station_hohenheim` [cited 2012-07-21].

Yankauer, A. (1990). Who are the peer reviewers and how much do they review? *Journal of the American Medical Association*, 263(10):1338–1340.

You, J. and Hubbard, K. G. (2006). Quality control of weather data during extreme events. *Journal of Atmospheric and Oceanic Technology*, 23(2):184–197.

Yuan, X. (2004). High-wind-speed evaluation in the southern ocean. *Journal of Geophysical Research*, 109(D13101):10.

Zahumensky, I. (2007). Guidelines on quality control procedures for data from automatic weather stations. World Meteorological Organization, WMO-No. 488, Appendix VI.2.

Acknowledgement

First of all I would like to thank Prof. Dr. Andreas Hense, for helping me with getting acquainted with the topic of statistics and scientific work. I thank him for his ideas and his advice and especially for allowing me so much freedom in my work in the past years.

I also thank the members of the project "Publikation Umweltdaten". Having three different views in this interdisciplinary project, allowed to work very fruitful in many ways. Especially Florian Quadt and Heinke Höck helped me a lot with understanding the problematics from their point of view. Our discussions and their help gave me a valuable insight into the ideas behind a data centre and computer science. Thanks also to Prof. Dr. Andreas V. Hense, Dr. Michael Lautenschlager and Dr. Martina Stockhause for the fruitful discussions during the project meetings. The bachelor and master students from the Bonn-Rhine-Sieg University of Applied Sciences in this project (Martin Dames, Christoph von Hinten, Marcel Sponer, Lukas Brabanski, Martin Müller, Benedikt Huber) must not be forgotten, since their support and their views helped the project to work very well.

I would further like to thank the members of the "Working Group on Climate Dynamics", for the discussions in- and outside the meetings. Thanks also to the other members of the meteorological institute, who allowed by their open-door policy a positive atmosphere for work and more. Especially, I would like to thank Elke Rustemeier for our many discussions on homogenisation and quality control. Thank you for giving me so much advice on literature and for making me feel, that my topic has at least a little bit to do with meteorology.

Further thanks go to those, who helped me with a great deal of ideas concerning the language corrections: Lilo Bach, Armin Bruß, Matthieu Masbou, Christoph Wosnitza and Adriana Wyss.

Furthermore, I would like to thank those individuals and groups, contributing to the R project (R Development Core Team [2011]), which allowed me to perform the calculations and the plotting in this thesis. Especially the following packages are used for the plots and calculations of this thesis: fields (Furrer et al. [2011]), ncdf (Pierce [2011]), qat (Düsterhus [2011]), car (Fox and Weisberg [2011]), nlme (Pinheiro et al. [2011]), boot (Canty and Ripley [2011]), XML (Lang [2011]), moments (Komsta and Novomestky [2011]), gdata (Warnes et al. [2011b]) and gplots (Warnes et al. [2011a]).

I would also like to thank those institutions and groups, who made the used data of this thesis available. First, the German National Weather Sevice (DWD), which provides some data for its climate network at http://www.dwd.de/klimadaten, is to mention. Secondly, the European Centre for Medium-Range Weather Forecasts (ECMWF), which made the data of the ERA40 reanalysis available. Then the National Oceanic and Atmospheric Administration (NOAA), which provides the data of the NCEP Reanalysis on their Web site at http://www.esrl.noaa.gov/psd/. I would also like to thank the contributors to the General Observation Period (GOP), especially the group of the Institute of Physics and Meteorology at the University of Hohenheim led by Prof. Dr. Volker Wulfmeyer and the group of Prof. Dr. Susanne Crewell of the Institute of Geophysics and Meteorology at the University of Cologne for their provision of the data of this project. In addition, I thank the Climate Research Unit of the University of East Anglia and the Met Office for the provision of the HadCRUT3 dataset and the basic data at http://www.metoffice.gov.uk/research/climate/climate-monitoring/land-and-atmosphere/surface-station-records. Therein, I would like to thank also all national meteorological services, which allowed the Met

Office to distribute the basic data of this temperature reconstruction.

Thanks a million to my family for all their support in the past years and for believing in me.
At most I would like to thank Lilo for helping me through the past years and for listening to all my complaints about this thesis and my work. For having a warm word in any situation, offering a different view and saving me from losing my head. Thanks!

"To the end, that such Productions being clearly and truly communicated, desires after solid and usefull knowledge may be further entertained, ingenious Endevours and Undertaking cherished, and those, addicted to and conversant in such matters, may be invited and encouraged to search, try and find out new things, impart their knowledge to one another, and contribute what they can do to the Grand design of improving Natural knowledge,[ ]and perfecting all Philosophical Arts, and Sciences. All for the Glory of God, [...] and the Universal Good of Mankind."
(from the Introduction of the Philosophical Transactions for the Royal Society (Philosophical Transaction Staff [1665]))