The analog-method as statistical upscaling tool for meteorological field reconstructions over Northern Europe since 1850

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
im Fachbereich Geowissenschaften
der Universität Hamburg

vorgelegt von

Frederik Schenk

aus

Leimen

Hamburg
2015

Als Dissertation angenommen vom Fachbereich Geowissenschaften der Universität Hamburg

auf Grund der Gutachten von Prof. Dr. Dr. h.c. Hans von Storch
und Dr. Eduardo Zorita

Hamburg, den 20.12.2013

Prof. Dr. Christian Betzler
Leiter des Fachbereichs Geowissenschaften

Abstract

In this study, the analog-method is introduced, evaluated and applied as a nonlinear statistical upscaling tool to reconstruct multivariate atmospheric fields from long historical station records. The difficulty of currently available high resolution meteorological fields is their shortness mostly limited to the last up to 60 years. In contrast, a growing number of homogenized daily pressure and monthly mean temperature observations allow to study regional climate variations over Northern Europe back to 1850. The aim of this study is hence to develop a method which combines the advantages of numerically downscaled atmospheric fields by a regional climate model (RCM) for the last 50 years with statistical upscaling of long homogenized pressure and temperature observations since 1850.

This study shows that this aim can be achieved by the analog-method. Based on a set of long historical station records, the method searches for the highest pattern similarity between every time step in the past and its closest "analogue" stored in an archive of observations of the last 50 years. In this archive, corresponding fields to the station records are available from the RCM which has been forced by reanalysis data at its boundaries. Based on the maximized similarity of the analogous pairs, the method assumes that also the corresponding atmospheric states should be very similar. Atmospheric fields since 1850 are then reconstructed by taking the RCM fields from its closest analog in the last 50 years.

The first part of the thesis focuses on testing different settings of the analog-method to evaluate the robustness of the field reconstruction on daily to monthly scale. This involves changes in the number and spatial distribution of predictor stations, the size of the analog-pool, the stationary of the calibration period, the density of suitable neighbouring analogs, frequency distributions and the introduction of persistence into the reconstruction of daily variables. As a measure of skill, temporal correlations, mean bias and variance are estimated between the reconstruction and the RCM simulation. The validation confirms a good reconstruction skill with realistic statistical properties for variables with a strong physical link to pressure and temperature.

In the second part, the reconstruction is extended to the period 1850-2009. Low-frequency variations and long-term trends of annual storminess are compared to the novel 20th century reanalysis (20CR) since 1871 and different pressure-based storm indices over Northern Europe. The analog-reconstruction shows a highly realistic reproduction of the storm climate since 1850 compared to storm indices derived from homogenized pressure observations. In contrast, large inconsistencies in 20CR prior to 1920-40 lead to spurious long-term trends i.e. over open sea areas. The reconstruction confirms that no long-term trend exists in annual storminess while the observed NE-shift of storm tracks since the 1970s is unprecedented since 1850. Large decadal variations and a return to average conditions in the last decade indicate no systematic change so far.

The third part of the thesis provides a thorough validation of the temperature reconstruction's ability to reproduce spatial coherency of the trend patterns, low-frequency variations and trends of seasonal temperatures since 1850. While seasonal correlations are high with very realistic long-term trends, the evaluation suggests also a dependency on deficiencies inherited from the used RCM. A larger analog-pool and spectrally nudged RCM fields should allow here further improvements.

A brief overview on applications in chapter 5 confirms that the reconstructed atmospheric forcing fields allow a realistic simulation of the Baltic Sea ecosystem since 1850. While the evolution of eutrophication is clearly dominated by anthropogenic nutrient discharge, a reasonable reproduction of salinity, oxygen and water temperatures indicates realistic meteorological forcing conditions provided by the reconstruction. This is confirmed by a high agreement of simulated and observed sea-ice extents and a short evaluation of reconstructed vs. observed weather conditions connected to a strong inflow event into the Baltic Sea.

This thesis concludes that analog-upscaling provides an optimal tool to reconstruct physically consistent high resolution meteorological fields over Northern Europe based on sparse station records linked to the output of a RCM. Further improvements with the current settings should be mainly achieved by more realistic RCM fields e.g. by using spectral nudging. The results of this study motivate further applications of analog-upscaling in the context of paleoclimate reconstructions e.g. by directly searching for analogs in the RCM space of paleoclimate simulations.

Zusammenfassung

In dieser Arbeit wird die Analog-Methode als nicht-lineares statistisches Upscaling Instrument eingeführt, evaluiert und angewendet, um multi-variate atmosphärische Felder aus langen historischen Stationsdaten zu rekonstruieren. Die Schwierigkeit von derzeit verfügbaren hoch aufgelösten meteorologischen Feldern ist deren zeitliche Begrenztheit auf die letzten bis zu 60 Jahre. Im Gegensatz dazu erlaubt eine wachsende Anzahl an homogenisierten täglichen Luftdruck- und monatlichen Temperaturmessdaten eine Analyse regionaler Klimavariabilität über Nordeuropa zurück bis 1850. Das Ziel dieser Studie ist daher die Entwicklung einer Methode, die die Vorteile numerischen Downscalings atmosphärischer Felder durch ein regionales Klimamodell (RCM) der letzten 50 Jahre mit statistischem Upscaling langer homogenisierte Druck- und Temperaturmessdaten seit 1850 verknüpft.

Diese Studie weist nach, dass das Ziel mit der Analog-Methode erreicht werden kann. Basierend auf einem Satz langer historischer Stationsdaten sucht die Methode nach der höchsten Ähnlichkeit der Muster zwischen jedem Zeitschritt in der Vergangenheit und seinem nächsten Analog, das in einem Archiv von Beobachtungsdaten der letzten 50 Jahre gespeichert ist. In diesem Archiv sind nun die korrespondierenden Felder zu den Stationsdaten verfügbar, die durch ein RCM bereitgestellt werden, welches an seinen äußeren Rändern mit Reanalysedaten angetrieben wurde. Die Analog-Methode geht nun davon aus, dass basierend auf der maximalen Ähnlichkeit der Analog-Paare auch die korrespondierenden atmosphärischen Felder sehr ähnlich sein sollten. Aufgrund dieser Annahme werden nun die Felder ab 1850 durch die RCM-Felder des ähnlichsten Analoges aus den letzten 50 Jahren rekonstruiert.

Der erste Teil der Arbeit konzentriert sich auf das Testen verschiedener Einstellungen der Analog-Methode um die Robustheit der Feldrekonstruktion auf täglicher und monatlicher Zeitskala zu evaluieren. Die Tests umfassen Änderungen der Anzahl und räumlichen Verteilung der als Prädiktor verwendeten Stationen, den Umfang des Analogpools, die Stationarität der Kalibrierungsperiode, die Dichte geeigneter benachbarter Analoge, Häufigkeitsverteilungen und die Einführung von Persistenz in die Rekonstruktion täglicher Variablen. Als Gütemaß werden die zeitliche Korrelation sowie die Abweichung der Mittelwerte und Varianz zwischen der Rekonstruktion und der RCM Simulation bestimmt. Für Variablen mit einer starken physikalischen Verbindung zu Luftdruck und Temperatur bestätigt die Validierung eine hohe Rekonstruktionsgüte inkl. realistischer statistischer Eigenschaften.

Im zweiten Teil der Arbeit wird die Rekonstruktion auf die Periode 1850-2009 ausgedehnt. Niederfrequente Variationen and langfristige Trends jährlicher Sturmstatistiken werden dann mit der neuen 20th Century Reanalysis (20CR) ab 1871 und verschiedenen luftdruckbasierten Sturmindices über Nordeuropa verglichen. Es zeigt sich, dass die Analogrekonstruktion das Sturmklima ab 1850 sehr realistisch reproduziert, wenn diese mit Sturmindices verglichen werden, die auf homogenisierten Luftdruckdaten basieren. Im Gegensatz dazu führen große Inkonsistenzen in 20CR vor etwa 1920 oder 1940 zu fehlerhaften Langzeittrends v.a. über offenen Meeresgebieten. Die Rekonstruktion bestätigt hingegen, dass keine Langzeittrends in jährlichen Sturmstatistiken existieren während aber die beobachtete Nordostverlagerung der Sturmbahnen seit den 1970ern beispiellos ist seit 1850. Große dekadische Schwankungen und

eine Rückkehr zu durchschnittlichen Bedingungen in der letzten Dekade zeigen zumindest bisher keine systematischen Änderungen an.

Der dritte Teil der Arbeit widmet sich einer gründlichen Validierung der Fähigkeit der Temperaturrekonstruktion räumlich kohärente Trendmuster, niederfrequente Schwankungen sowie Trends saisonaler Temperaturen seit 1850 zu reproduzieren. Während die saisonalen Korrelationen hoch sind mit gleichzeitig sehr realistischen Langzeittrends, deutet die Auswertung auch auf Fehler hin, die durch die RCM-Felder vererbt werden. Ein größerer Analogpool und spektral genudgte RCM-Felder sollten hier weitere Verbesserungen ermöglichen.

Ein kurzer Überblick zu Anwendungen der Rekonstruktion in Kapitel 5 bestätigt, dass die rekonstruierten atmosphärischen Antriebsfelder eine realistische Simulation des Ökosystems der Ostsee ab 1850 erlauben. Während die Entwicklung der Eutrophierung klar durch menschliche Nährstoffeinträge dominiert wird, deutet die angemessene Wiedergabe von Salzgehalt, Sauerstoff und Wassertemperaturen auf realistische meteorologische Antriebsdaten durch die Rekonstruktion hin. Dies wird durch eine hohe Übereinstimmung der simulierten und beobachteten Eisbedeckung und durch eine kurze Auswertung von rekonstruierten Wetterbedingungen mit Bezug zu einem starken Einströmungsereignis in die Ostsee im Vergleich zu Beobachtungen bestätigt.

Durch diese Arbeit kann bestätigt werden, dass Analog-Upscaling ein optimales Instrument darstellt, um physikalisch konsistente hoch aufgelöste meteorologische Felder über Nordeuropa auf Basis spärlicher Stationsdaten zu rekonstruieren, welche mit RCM-Feldern verknüpft sind. Weitere Verbesserungen mit den aktuellen Einstellungen sollten hauptsächlich durch realistischere RCM-Felder z.B. durch Verwendung von spektralem Nudging erreicht werden. Die Ergebnisse motivieren eine erweiterte Anwendung von Analog-Upscaling im Kontext von Paläoklimarekonstruktionen z.B. durch eine direkte Suche nach Analogen in Feldern einer Paläoklimasimulation.

Content

1 Introduction

1.1 Motivation

The global climate of the last 150 years is characterized by a strong warming trend with different impacts on regional climates and ecosystems. While the northern hemisphere has warmed by +0.76 K per century during the period 1861-2010 (Jones et al., 2012), annual warming trends over the northern Baltic Sea region reached rates up to +1.1 K per century from 1871 to 2011. Besides the clear temperature change in this period, the Baltic Sea ecosystem has undergone also large changes due to human activities in surrounding countries which led to increasing eutrophication i.e. since the middle of the last century. While a quantification of the impact of climate change on the Baltic Sea ecosystem in recent decades is not clear, human impacts through nutrient overload and fishery are currently a dominant factor posing a growing threat onto a healthy state of the Baltic Sea ecosystem.

The Baltic Sea currently holds the world's largest dead zone caused by anthropogenic influence with a hypoxic (oxygen depleted) area of about 60.000 km² per year in last decades (Conley, 2012). While dead zones occur in many places all over the world, the situation of the Baltic Sea as a semi-enclosed marginal sea basin surrounded by densely populated countries requires special scientific and political attention. Although periods of increased hypoxia have occurred in the Baltic Sea throughout the Holocene (Zillén et al., 2008; Zillén and Conley, 2010), the combination of anthropogenic nutrient discharge with warming trends have led to political consequences defining national nutrient load reductions through the Baltic Sea Action Plan by the Helsinki Commission (cf. HELCOM, 2011). Although the action plan was adopted in 2007 and nutrient loads decreased since around 1980, no improvement of the water quality has been achieved so far. Plans to use geoengineering e.g. through pumping oxygen into deep oxygen depleted basins of the Baltic Sea are seen critically due to the high costs and potential unforeseeable impacts on the ecosystem (Conley, 2012).

For a better understanding and analysis of ongoing and potential future changes of the Baltic Sea ecosystem, one aim of the project ECOSUPPORT (Advanced tool for scenarios of the Baltic ECOsystem to SUPPORT decision making) from 2009 to 2011 (Meier et al., 2012a, 2014) was to develop a multi-model system tool to simulate relevant processes reaching from regional climate models over different marine ecosystem, food and fish models to economic calculations. To estimate the skill and uncertainties of the models and to study the timescales of involved processes like eutrophication, also long historical simulations from the pristine state of the Baltic Sea around 1850-70 until today were needed. As the required multivariate meteorological forcing fields for such ecosystem simulations have not been available on longer timescales, the aim of this thesis is to develop and apply a new reconstruction method which is able to reconstruct High Resolution Atmospheric Forcing Fields (HiResAFF) since 1850 over Northern Europe based on a set of homogenized historical station data of daily sea-level-pressure and monthly mean near-surface temperatures.

In the following, the needed properties of the reconstruction are identified for the reconstruction design. Based on these criteria, the analog-method is introduced, evaluated and applied as statistical upscaling tool to derive the required fields of HiResAFF in chapter 2. Chapter 3 and 4 aim at analysing the long-term variations and trends of storminess and seasonal temperatures in comparison to different observations since 1850. Chapter 5 presents some applications of HiResAFF used to drive ecosystem models of the Baltic Sea. The main results regarding the properties and robustness of the method are summarized in chapter 6.

1.2 Requirements for meteorological forcing fields

To realistically drive the dynamics and air-(ice-)-sea-fluxes of complex ocean- or marine ecosystem models used in this project (Meier et al., 2012b), meteorological forcing fields of sea-level-pressure, near-surface wind speeds (U, V), precipitation, humidity, cloud-cover and near-surface temperatures are needed. Due to the topographic situation of the Baltic Sea with its complex bathymetry, different sub-basins connected over shallow sills and only narrow connections to the North Sea over the Belts, a very high resolution is required for the models and the forcing fields to realistically reproduce the regional details of the marine circulation which dominate local to basin wide processes.

Global reanalysis data

The required fields with high spatial resolution and physically consistent properties are ideally produced by regional climate models (RCM) numerically downscaling coarsely resolved global reanalysis data to the local scale. Global reanalysis datasets like NCEP/NCAR (Kistler et al., 2001) are however limited to the period since 1948. Only recently, the 20th Century Reanalysis (20CR, Compo et al., 2011) became available covering the period since 1871. Different to NCEP, only daily pressure, monthly mean sea-surface-temperatures (SST) and monthly ice distributions are assimilated into the global climate model of 20CR. By assimilating all pressure and SST data available, 20CR produces the best dynamical solution for the atmospheric state of a given time step.

However, there is clear evidence already for the shorter reanalysis of NCEP that spatial and temporal changes in the number, type or quality of assimilated observations can introduce spurious trends or offsets into the simulation (Kistler et al., 2001, Bengtsson et al., 2004). This is most pronounced over sea areas of the southern hemisphere where little observations exist back in time i.e. before the satellite era 1979 (cf. Hines et al., 2000, Bromwitch et al., 2007). First studies indicate that also 20CR is compromised by a reduced reliability within the last 50 years over the southern hemisphere (Compo et al., 2011). At least on regional scales clear inconsistencies have been identified on longer timescales in storminess over the NE-Atlantic and North Sea (Krueger et al., 2013; chapter 3) or temperature and precipitation over the US (Fergusson and Villarini, 2012) making 20CR unsuitable for long-term analysis.

Gridded observational fields

Large efforts have been undertaken to produce gridded datasets of different variables from observations. The fields are however not available for all variables on longer timescales and/or provide insufficient properties to realistically drive ocean- or ecosystem models. As an example, the newest high resolution gridded temperature datasets of CRU TS 3.10 (Harris et al., 2013) with a horizontal resolution of 0.5°x0.5° since 1901 does not provide temperature conditions over sea areas of the Baltic Sea. The gridded dataset of CRUTEM4 (Jones et al., 2012) provides temperatures since 1850 with a horizontal resolution of 5°x5° so that terrestrial temperatures would be used as forcing also over Baltic Sea areas. As little to no air temperature or wind records exist over sea areas back in time, it is difficult to achieve realistic surface conditions over sea areas directly from measurements.

1.3 The concept of field reconstruction through analog-upscaling

Due to the shortness or long-term inconsistencies of reanalysis data and the very coarse resolution of gridded observations on longer timescales, the required meteorological forcing fields need to be reconstructed from available homogeneous station data covering Northern Europe at least since 1870 to 1850. The problem to predict the atmospheric fields of different state variables based on only some local records was already a well known problem in the context of short-term weather forecasting in the 1970s.

In this context, Lorenz (1969) "invented" the concept of the analog-method in the late 1970s when numerical weather prediction models (NWP) were not yet in common use i.e. due to the low computing power available at that time. The idea was to search for a most similar analog for the current atmospheric state in the past. Then it was assumed that the following evolution in the past can serve as a forecast of the current situation. The predictive skill was however very low. First, there were too little observations to always find suitable analogs (van den Dool, 1994). Second, very small differences of the initial analog lead to large subsequent deviations in the forecast – a problem also well known for today's NWP. As the initial value problem in a chaotic system is less serious on longer timescales, the analog-method was slightly more successful for long-range weather forecasting (e.g. Toth, 1989) or short-term climate predictions (e.g. Barnett and Peisendorfer, 1978, van den Dool, 1994).

Independent from these limitations and a poor predictive skill, Toth (1989) describes the main advantage of analog-forecasting *"…that it yields a real solution to a difficult problem. It automatically contains the microclimatological influences at every locations of the forecast."* This aspect is picked up in this thesis for the reconstruction of atmospheric fields from sparse observations. As described in more detail in chapter 2, the principal idea is to first use long historical station records to find for any day/month etc. in the past a very similar spatial pattern for the same stations e.g. within the last 50 years. For the latter period, the multivariate fields to the station records are well known from observations e.g. provided by a RCM which has been driven by reanalysis data ("observations") at its lateral boundaries. The station records in the 50 year archive are hence temporally correlated with the corresponding RCM fields. The basic assumption is now that when a set of station records of a target day/month etc. in the past is very similar to its closest analog of the same stations in the 50 year archive, then also the corresponding atmospheric fields should be very similar. The solution to a difficult problem (Toth, 1989) is hence to reconstruct the unknown atmospheric fields in the past by the known fields from its closest analog of the last 50 years provided by the RCM.

1.4 Aim and structure of this study

The concept of analog-upscaling described here can be understood as a nonlinear empirical transfer function which links a set of long historical station data to multivariate atmospheric fields of a usually shorter period e.g. provided by a RCM. The assumption that the pattern similarity between analogs in the predictor space (station records or proxies) should lead to a reasonable similarity also between the target field and the field of the analog should be valid if the predictor variable (e.g. daily pressure records) has a meaningful link to the variable of the target fields (e.g. temperature, wind, precipitation etc.).

To test whether the analog-reconstruction has any meaningful skill, the main task of this thesis is to falsify the null-hypothesis (H_0) that the temporal correlation (the similarity) between the analog-reconstruction and reference fields is not significantly different from zero. Hence, it has to be shown that the redistribution of the RCM fields according to the pattern

similarity found within the predictor space yields a significantly higher correlation (A_1) than just randomly choosing the RCM fields. As mentioned above, the question whether H_0 (no skill) is wrong and A_1 (significant skill) is true, will mainly depend on the chosen predictor variable and the physical link to the variables of the target fields and also on the definition of the similarity measure for the predictor space. As explained in more detail in chapter 2, station records of daily pressure and monthly mean temperature are used to reconstruct multivariate meteorological fields and the Euclidian distance will be used as similarity measure to find the closest analogs within the predictor space.

In chapter 2, the standard settings for the analog-reconstruction are defined and different settings are tested to evaluate the robustness of the reconstruction. The latter is corroborated through the dependency of the reconstruction skill on the number and spatial distribution of the predictors, the size of the analog-pool, the stationary of the calibration period and the density of suitable analogs for different variables of the target fields. As a measure of skill, temporal correlations of the reconstruction relative to a reference simulation are used on daily and monthly scale. For the reconstruction of HiResAFF, the reconstruction skill regarding mean bias, daily and monthly variance and the agreement of frequency distributions relative to the reference simulation of the RCM will be evaluated.

Because the analog-method searches for the best analog independently from the neighbouring analog, the time-invariant reconstruction partly destroys the serial correlation typically found for most daily geophysical time series. It is briefly assessed to which extent autocorrelations are realistically reconstructed and how persistence can be introduced to the analog-approach.

As daily pressure has a weak physical link to predict long-term variations and trends in temperature, an alternative reconstruction is developed in chapter 2 through predicting the monthly mean temperature fields by monthly mean station records while daily temperature anomalies are predicted by daily pressure stations. As the reconstruction of monthly analogs considerably reduces the amount of available analogs compared to daily scale, chapter 4 analysis whether this approach is sufficient to reproduce long-term seasonal variations and trends in comparison to observations since 1850. In this context, it is also tested to which extent spatial trend patterns are realistically reproduced by the analog-method and to which extent spatial correlations relative to observations dependent on the used RCM fields.

In chapter 3, the pressure-based daily reconstruction of HiResAFF is extended to the period 1850-2009. The focus is here on the analysis of long-term variations and trends of annual storminess since 1850 over Northern Europe and its consistency with other long-term storm information. The results of HiResAFF are compared with those derived from the novel 20CR since 1871. To decide whether inconsistencies in low-frequency storm variations and trends between both datasets are caused by the reconstruction or 20CR, station pressure-based storm indices are used as an independent source for the validation of HiResAFF and 20CR. As recently shown, these pressure-based storm indices provide mostly a fair to high informational content to study long-term changes in wind statistics (Krueger and von Storch, 2011; 2012). A thorough cross-comparison is applied to achieve a more conclusive view on the reliability of the different storminess measures and to answer the question whether any changes can be detected in the storm climate of Northern Europe since around 1850.

Within the ECOSUPPORT project, the multivariate reconstruction of HiResAFF has been applied to force three different ecosystem models of the Baltic Sea. Chapter 5 gives a brief overview about some results of the so far longest transient simulations of the Baltic Sea state since 1850. Some examples are given to demonstrate whether reconstructed high-frequency

variations in pressure and wind and low-frequency variations and trends in seasonal temperatures of HiResAFF have been of suitable quality to realistically reproduce bio(geo)chemical and physical properties of the Baltic Sea on long timescales in comparison to marine observations since 1860.

Chapter 6 finally summarizes the main results and conclusions about the strengths and weaknesses of analog-upscaling for different variables and properties on daily to centennial scale. As the current thesis has only evaluated some aspects of the method, some possibilities to improve high-frequency correlations and achieve more analogs for potential long-term reconstruction in the context of paleoclimate are motivated.

2 Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling

The analog method (AM) has found application to reconstruct gridded climate fields from the information provided by proxy data and climate model simulations. Here, we test the skill of different setups of the AM, in a controlled but realistic situation, by analysing several statistical properties of reconstructed daily high-resolution atmospheric fields for Northern Europe for a 50-yr period. In this application, station observations of sea-level pressure and air temperature are combined with atmospheric fields from a 50-yr high resolution regional climate simulation. This reconstruction aims at providing homogeneous and physically consistent atmospheric fields with daily resolution suitable to drive high resolution ocean and ecosystem models.

Different settings of the AM are evaluated in this study for the period 1958–2007 to estimate the robustness of the reconstruction and its ability to replicate high and low-frequency variability, realistic probability distributions and extremes of different meteorological variables. It is shown that the AM can realistically reconstruct variables with a strong physical link to daily sea-level pressure on both a daily and monthly scale. However, to reconstruct low-frequency decadal and longer temperature variations, additional monthly mean station temperature as predictor is required. Our results suggest that the AM is a suitable upscaling tool to predict daily fields taken from regional climate simulations based on sparse historical station data.

2.1 Introduction

The availability of gridded meteorological forcing data is a prerequisite for many climate impact related studies including hydrological, ocean or ecosystem simulations. Detection and attribution studies, e.g. for the climate of Baltic Sea catchment (Bhend and von Storch, 2008; 2009), are typical research topics where recent potentially anthropogenic changes in the climate system need to be detected by comparing them to the natural climate variability undisturbed by human impacts. While such studies can be done based on coarsely resolved gridded data of single variables, the detection and attribution of environmental changes including eutrophication, e.g. within the Baltic Sea ecosystem, require a full set of meteorological variables to force related bio(geo)chemical models (cf. Meier et al., 2011a; 2012b; 2014; Gustafsson et al., 2012 and chapter 5).

State-of-the-art regional climate models (RCM) are a common tool to provide such highly resolved and physically consistent atmospheric fields for a given domain by numerically downscaling global reanalysis data, for instance related to NCEP/NCAR-reanalysis since 1948 (Kistler et al., 2001), ERA40-reanalysis since 1957 (Uppala et al., 2006) and ERA-Interim (Dee et al., 2011) since 1979. However, longer simulations spanning the whole 20th century or even longer would allow estimating the longer-term variability including periods in which the anthropogenic greenhouse gas forcing was not as strong as in the last few decades.

One possibility to reconstruct high-resolution meteorological fields is to conduct simulations with a RCM driven at the boundaries by global general circulation models (GCM) over the past decades or few centuries (cf. PRUDENCE project; Vidale et al., 2003; Giorgi et al., 2004; Déqué et al., 2005). Although regional simulations of the past millennium, e.g. over the Baltic Sea (Graham et al., 2009; Schimanke et al., 2012), provide an important test bed to study the impact of external forcing on the regional climate, the time evolution of the

simulated meteorological fields is not guaranteed to be close to the evolution of the real meteorological fields because, in the absence of data assimilation, the internal variability of the model and observations will in general be uncorrelated in time. In addition, the model bias introduced by the GCM-RCM simulations usually leads to considerable (systematic) deviations from the observed climate even if ensembles of different models are used (Jun et al., 2008). Due to the deviation in time between GCM runs and observations, statistical downscaling as another approach to bridge the gap between the coarse resolution of the large-scale data of the GCM and the regional or local state of the atmosphere (von Storch et al., 1993; Zorita and von Storch, 1999; Frías et al., 2006; Matulla, 2005) cannot be used here. In addition to errors/uncertainties introduced by the statistical model, it is difficult to estimate if the relationship established is also valid outside the reference period on longer time scales, i.e. if the process linking the large-scale with the local scale is stationary (cf. Bürger et al., 2006 for regression).

As several long station observations are available for Northern Europe reaching back to 1850, statistical upscaling provides another possibility to reconstruct atmospheric fields. This can be done either by different interpolation techniques or by setting up an empirical relation between observations and the large-scale atmospheric field. The general difficulty of this approach is to reconstruct atmospheric fields with high spatial resolution from a limited number of observations. To achieve physically consistent atmospheric fields with realistic probability distributions, we set out to develop an upscaling tool that combines the information provided by long time series from a few stations together with simulations from RCMs with high spatial resolution that only span the rather short reanalysis period. The statistical method applied here to combine both sources of information – and in the end provide full high-resolution meteorological fields over a longer period than that spanned by the reanalysis – is the analog method (AM).

The AM is a kind of non-linear empirical transfer function that allows one to estimate a set of predictands from a set of known predictors. Usually, the sets of predictors cover a longer time span than the predictands and the AM aims at estimating the predictands in the period where they are not available, based on the information provided by the predictors. The basic idea of the AM itself was introduced earlier into the field of weather prediction in the late 1970s (cf. Lorenz, 1969; Kruizinga and Murphy, 1983), followed by studies of short-term climate prediction (Barnett and Peisendorfer, 1978; van den Dool, 1994).

The general idea is illustrated in Fig. 2-1. Denoting the time step t for which an estimation of the predictands is needed, the AM searches through a data archive P(u) in which predictor P(t) and predictand P(u) are both available, and identifies the time step u in which the predictor is closest to its value at time t, its analogue. The imputed predictand for time t is then the value of the predictand at time u. Variants of the AM can be introduced by defining different measures of similarities between the predictor at time t and at time u to weight more strongly some properties of the analogues that might be desired for particular applications. Other possibilities lie in augmenting the time window around time steps t and u, thereby searching for analogous successions, instead of just analogous snapshots, to retain the serial correlation that may be present in the predictand (Fig. 2-10).

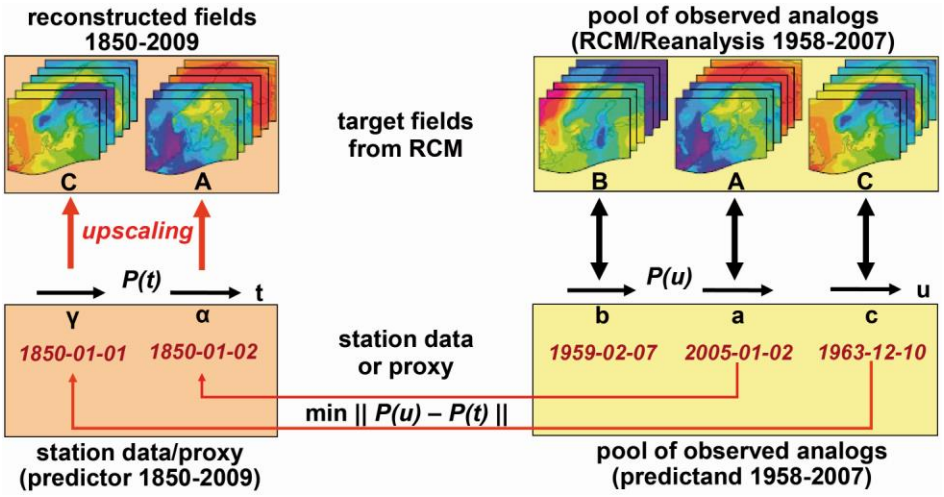

Fig. 2-1. Scheme for the Analog-Method used as upscaling tool. Any day a, b, c, etc. of P(u) is linked to its related fields A, B, C, etc. taken from RCM/Reanalysis (predictand, analogs). The fields of a historical day γ, α etc. from P(t) is found by Eq. (1) to be most similar to c, a, etc. in P(u) forming the analog-pool. Hence, it is assumed that the fields of γ, α etc. are then very similar to the fields C, A, etc. (upscaling).

The AM requires a data archive that is large enough for sufficiently close analogues to be found. This size increases with the number of degrees of freedom required to specify the predictor. If the analogue space is one dimensional, it is relatively easy to find a close enough analog if the range of variability is not very different through time. If, however, the predictor is a multidimensional field with a large number of degrees of freedom, it will be generally difficult to find an analog that is close to the target along all dimensions. In this case, a very large data archive is required (van den Dool, 1994).

The AM was applied to climate research by Zorita et al. (1995). Following this approach, Cubasch et al. (1996) and Biau et al. (1999) empirically downscaled GCM output to the regional scale with the AM. It was found that the AM performs as well as more complicated empirical downscaling methods (Zorita and von Storch, 1999). Encouraged by these findings, the AM was further evaluated by Fernández and Sáenz (2003) who evaluated the analog search in predictor fields whose dimensionality had been previously truncated by either the classical principal component analysis (PCA) or by a canonical correlation analysis (CCA).

Matulla et al. (2004) applied the AM for the estimation of local temperature and precipitation change scenarios at daily scale over complex terrain like Austria (thus a downscaling application of the AM). In their study, the authors highlight the importance of choosing the appropriate predictor variables to obtain meaningful physical links to the local predictand. As an example, they show that changes in sea-level-pressure in a global scenario would fail to describe the warming produced by anthropogenic greenhouse gas forcing, whereas the skill of sea-level-pressure to capture precipitation changes is high. On the other hand, using additionally relative topography as predictor, local warming is well predicted, whereas precipitation is strongly underestimated.

Wetterhall et al. (2005) evaluated the AM as benchmark method for downscaling precipitation over Central Sweden, e.g. dependent on the domain size and different similarity measures. In a more recent study, Matulla et al. (2008) evaluated whether other similarity measures than the commonly used Euclidian distance for the AM are better suited for downscaling daily precipitation from the large scale circulation provided by a GCM. They concluded that the

Euclidian distance performs better or at least as well as more complicated similarity measures. In addition, they showed that, when searching for analogous successions, a stronger weighting of the previous three days of a precipitation event can improve the skill of the AM. The performance of the AM in general increases with increasing precipitation amount; hence the AM has difficulties in accurately reproducing low rainfall or dry days. In contrast, the AM shows good performance in estimating dry local scale conditions from large-scale circulation on monthly timescales.

To our knowledge, the AM has only recently been used in climate research as a statistical upscaling tool in the framework of paleoclimate. Graham et al. (2007) formally introduced the AM as a "proxy surrogate reconstruction" (PSR) using atmospheric fields from a coupled Atmosphere Ocean General Circulation Model (AOGCM) as predictand and a set of proxy records as predictors to reconstruct the full global meteorological fields compatible with the information provided by the proxy records. The same technique was also used by Trouet et al. (2009), applying the PSR as a "proxy model analog method" to reconstruct the North Atlantic Oscillation (NAO) since 1000 AD from multi-proxy records by re-ordering the most similar surrogates from an AOGCM. The basic idea of the AM was also used in a different approach by Guiot et al. (2010), applying a "spectral analog method" as one part of a sophisticated model chain to reconstruct European temperatures back to 600 AD from different proxies. Similar to the idea of Moberg et al. (2005), different proxies were used by Guiot et al. (2010) to reconstruct different signals by splitting the proxies into three frequency bands to account for low (lake or ocean sediments), mid and high frequency (i.e. tree-rings) variations.

The principal advantage of the AM compared to regression methods has been shown by Fernández and Sáenz (2003). Although linear empirical downscaling methods perform as well as the non-linear AM downscaling approach when they are benchmarked by the correlation between reconstruction and target, they fail to reproduce a realistic variance and the non-normal distribution e.g. for daily precipitation is partly lost. The same problem is also typical for upscaling methods based on linear regression which often strongly underestimate the variability of the predictand. This problem is caused by the presence of noise in the predictors (von Storch et al., 2004). Also, whereas the predictand reconstructed by a linear method is bound to have the same probability distribution as the predictor, the general advantage of the AM is that no assumption about the probability distribution of the data is necessary. Hence, it can be applied to predictors and predictands with different probability distributions without any intermediate transformation of variables. Furthermore, the reconstruction shows no loss in variance and preserves the spatial covariance in the predictand fields (Zorita and von Storch, 1999). One disadvantage of the AM is that, in contrast to linear regression methods, the reconstructions based on the AM cannot exceed the range of already observed atmospheric states, i.e. it cannot extrapolate to unprecedented states of a possibly strongly different past or future climate. In the daily reconstruction case, also singular extreme events, e.g. the atmospheric conditions leading to the severe storm flood in 1872 at the SW Baltic Sea (Rosenhagen and Bork, 2009), cannot be reconstructed if analogues are not present in the archive of predictands.

In this study, the AM used as non-linear upscaling tool is applied and evaluated to reconstruct High RESolution Atmospheric Forcing Fields (HiResAFF) based on a limited set of station data used as predictors. We restrict the used stations to those spanning more than 150 yr to anticipate the further application of the analog-reconstruction back to 1850 (chapter 3, 4 and 5; Meier et al., 2011a, 2012b, 2014; Gustafsson et al., 2012) where only a limited set of homogeneous data is available. The predictands have been generated by a high-resolution regional climate simulation driven at the boundaries by global meteorological reanalysis and

thus the simulated fields are co-related in time with the station data. As the model can evolve more freely in the interior of the model domain, the temporal agreement with observation is, however, not perfect. While some RCM use spectral nudging (e.g. von Storch et al., 2000; Yoshimura and Kanamitsu, 2008) to bring the simulation closer to observations, no such simulation is used in this study. The results presented here are therefore to some extent also model dependent and provide a conservative validation of the AM upscaling.

The structure of the paper is as follows: Sect. 2 presents the data and methods used in the study. Different test cases for the evaluation and statistical methods for validation are introduced. Section 3 presents the results of different test cases related to the robustness of the AM and a validation of the reconstruction for the period 1958–2007. The results are discussed in Sect. 4 before a summary and outlook on a further application of the AM on longer time-scales is given in Sect. 5.

2.2 Data and methods

2.2.1 Historical station data (predictor)

As daily predictor, historical station data of up to 23 stations providing daily sea-level-pressure (SLP) for Northern Europe (71° N to 48° N, 5° W to 37° E, Fig. 2-2) are used. Only stations providing at least 100 yr of data are considered in this study. Daily mean SLP data since 1850 is provided by the EMULATE project (Ansell et al., 2006). Gaps have been filled in and the record completed until 2009 by including data from ECA&D (European Climate Assessment & Dataset) (Klein Tank et al., 2002) and different research institutions. For estimates of the data quality of daily SLP we refer to Ansell et al. (2006). Data for filling gaps in the EMULATE stations and the updates have been compared in overlapping periods. All time series were checked for outliers and systematic break changes. Mountain stations in the southern domain partly failed to pass this test due to changes in the hypsometric reduction of station pressure to sea-level. In this case, we additionally compared these stations with neighbouring stations with lower elevation. Due to partly missing or inconsistent conversions of station pressure to sea-level, the whole affected periods, rather than just single days, were set to missing values. Additional stations with higher elevations were omitted because daily errors are considerably large, with deviations of several hPa.

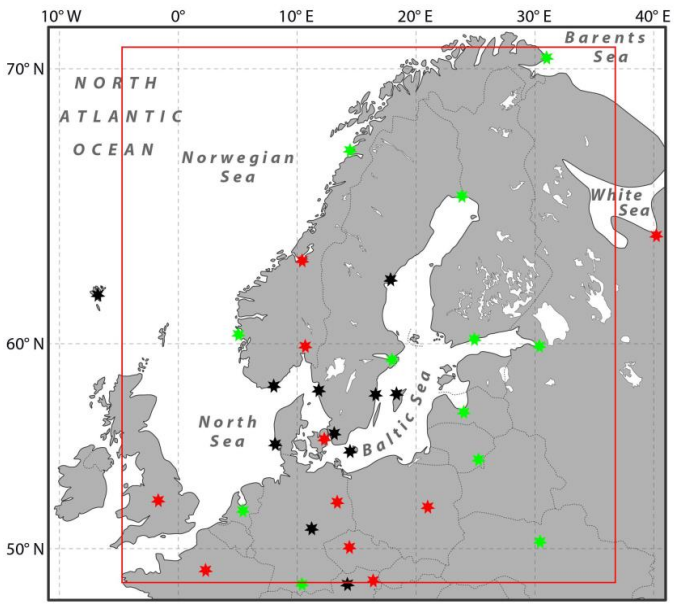

Fig. 2-2. Geographic positions of stations used as predictor in this study.

Green stations provide daily SLP and monthly T2M, black stations only daily SLP and red stations only monthly T2M. The domain of the reconstruction is indicated by the red rectangle.

More detailed information on the data and station location can be found in Tab. 3-1 and Tab. 4-1 for pressure and temperature, respectively.

Data coverage of daily SLP is partly lower in the 1950s and after 1990. As a consequence, the reconstruction during these periods shows a higher uncertainty due to a reduced number of predictors. The effect of a reduced number of predictors with different spatial distributions is elaborated in greater detail in Sect. 2.3.2.3. For a further extension of the reconstruction, we also include a test case where only six stations are used, as it would be the case when reconstructing atmospheric fields in 1850. It should be noted that the availability of more (sub-)daily data is still an ongoing work with still large differences among the several countries involved in these projects (Brunet and Jones, 2011) so that more predictor data will become available in the future.

At daily and monthly time scales, temperature variations at mid-and high latitudes are linked to the atmospheric circulation. In particular, winter temperatures in Northern Europe are known to be strongly modulated by the North Atlantic Oscillation (cf. Hurrel, 1995; Wanner et al., 2001). However, at longer time scales other factors like external climate forcing, greenhouse gases, aerosols, land use, etc. may play a stronger role. Thus, long-term trends in SLP do not necessarily evolve in parallel to long-term trends of temperature or other variables (Vautard and Yiou, 2009). Therefore, to capture the (multi-)decadal evolution of air temperature, monthly mean temperatures are reconstructed separately using monthly station temperature (T2M) as predictor (Sect. 2.2.3.4). For temperature, only 22 stations are selected from Jones and Moberg (2003) and Auer et al. (2007), which provide more than 100 yr of homogeneous data (Fig. 2-2). Whenever possible, the data were updated from the WMO database, ECA&D (Klein Tank et al., 2002) and the German Weather Service (DWD).

2.2.2 Atmospheric fields (predictand, analogs)

Focusing on Northern Europe and the Baltic Sea region, forcing fields with high spatiotemporal resolution are required in order to capture the high complexity of Baltic Sea sub basins. Therefore, multivariate atmospheric fields of mean sea-level pressure (SLP), 10 m wind (U, V), relative humidity (RELHUM), total cloud cover (TCLOUD), near-surface temperature (T2M) and precipitation (PREC) are taken from a climate simulation with the coupled Swedish Rossby Centre Regional Climate Atmosphere Ocean Model (RCAO; Döscher et al., 2002) over the last decades. The RCAO is used to numerically downscale ERA40 reanalysis data to a horizontal resolution of 0.25°×0.25° (~25 km) over Northern Europe for the period 1958–2007 (Meier et al., 2011b). Due to shortcomings in heat fluxes possibly related to the sea-ice model in RCAO, fields for the mean monthly temperature fields are taken from an atmosphere-only simulation with RCA3 without ocean (Samuelsson et al., 2011) that was additionally driven by observed sea-surface temperatures (Christensen et al., 2010). The output of the simulation is interpolated onto a regular geographical grid.

2.2.3 Methods

2.2.3.1 Basic concept of the analog-method as statistical upscaling tool

The AM assumes that given a spatial pattern of pressure, temperature or precipitation, etc. as target, it is possible to find a similar pattern in a set of observations. As illustrated in Fig. 2-1, denoting P as the vector of daily SLP observed in six stations on 1 January 1850, the AM compares it to all daily SLP patterns observed in January (or possibly winter months) during the period 1958–2007 (n = 1550 days) P(u) used as analog-pool. The day (e.g. 10 December 1963) for which the SLP is most similar to the target pattern is taken as the analog of 1 January 1850. In mathematical terms, the AM simply minimizes the distance between P and P(u) for each u from 1958–2007, the Euclidian distance:

$$\min \| P(u) - P(t) \| \quad (1)$$

In general, for each target pattern in the reconstruction period P(t) an analog P(u(t)) is found in the calibration period, based on the similarities of the corresponding SLP patterns. This analog mapping can be used to reconstruct fields of other variables, different to SLP and only available during the calibration period. In our previous example, the unobserved temperature or precipitation on 1 January 1850 is assumed by the AM to be very similar to the ones observed on 10 December 1963. This assumption will be valid if the predictor, SLP in this case, is strongly associated with the predictands (temperature or precipitation, etc.).

Here we have chosen the Euclidian distance in Eq. (1), but other similarity measures (Wetterhall et al., 2005; Matulla et al., 2008) may be chosen, for instance if the data recorded by some subset of the stations are assumed to be more accurate than the others. In general, there exist no optimal settings for the AM. They always depend on the particular purpose, i.e. which variables and which statistical properties of the reconstruction are of main interest. In practice, these subjective criteria for an optimal reconstruction should be defined previously before adjusting the AM accordingly. In this study, the optimal setting aims at reconstructing physically consistent fields of different variables based on the chosen predictor. This means that the settings used for the AM are not modified to optimize the reconstruction of each meteorological variable differently. Hence, the suggested higher weighting of previous days in the analog-search to improve the skill for reconstructed precipitation (Matulla et al., 2008) is not used for the sake of consistency with the other variables. Only in the temperature case, a modified approach needs to be used (Sect. 2.2.3.4).

2.2.3.2 Standard settings and application of the analog-method

Here, the standard setting for the AM applied for HiResAFF includes the use of the full analog-pool defined by the time span of 50 yr covered by the RCAO simulation driven by ERA40. The analog-pool consists of daily SLP predictor data for the period 1 January 1958 until 30 September 2007. In this period, the corresponding atmospheric fields (predictands) are also available from the simulation (Fig. 2-1). The daily reconstruction is separately produced for each of the twelve months of the annual cycle and possible analogs of a day in a given month m are searched in the month m and in the two months straddling m in the analog pool (M3 = m-1, m, m+1). This considerably increases the size of the analog pool and allows to reconstruct possible seasonal shifts through time. In the period covered by the analog pool, a day is reconstructed using a leave-one-out approach, i.e. the year of the target day is excluded in the analog search (otherwise the simulation would be exactly reproduced). Using the analog pool with M3 spanning over 50 yr yields around 4500 possible analogs for every historical target day in the reconstruction. The general effect of using smaller analog pools taken from different periods is evaluated in Sect. 2.3.2.1. Different settings are, however, required for the reconstruction of T2M (see below).

2.2.3.3 Implementation of persistence in the analog-method

In general, daily geophysical time series will display a serial correlation. As the AM in the standard approach searches the best analog for a defined target day, Eq. (1) does not explicitly optimize the search of the analog to replicate the dependence between consecutive days. However, serial correlation can still be implicitly captured by the AM if the serial correlation in the predictands is physically linked to the serial correlation present in the predictors (Fig. 2-10). Generally, the serial correlation of any variable in the predictands will be caused by several mechanisms, and it cannot be expected that the predictor captures them all. For instance, precipitation on day d may in general depend on precipitation on the previous day (e.g. accumulated soil moisture in summer) and not only on SLP on day d. As a consequence, the fields reconstructed by the standard AM setting will tend to display a weaker serial correlation than the original fields. However, persistence can be additionally implemented in

the AM. In order to consider the persistence in daily temperature predicted by daily SLP, Eq. (1) can be modified to search for the most similar sequence of n-lag days prior to P(t) including P(t). This means that an analog has to be found now in a space of dimension (nlag+1)·23. How many days (n-lags) are optimal for a realistic reconstruction of T2M persistence depends on the particular application and on the relevance of capturing the persistence in the predictand (Sect. 2.3.3.5).

2.2.3.4 Temperature reconstruction

Since daily SLP may be only weakly connected to temperature on longer time scales, the T2M fields have been reconstructed separately using information from station temperature data. Given that only monthly T2M station data are available prior to 1900, we split up the reconstruction of high-frequency and long-term temperature variations using different predictors. Daily temperature anomalies are reconstructed using daily SLP as predictor. The analog search is restricted to the month m = M1 of the target day because it is not possible to distinguish differences in the seasonal cycle of T2M based on daily SLP. Also, persistence is captured searching the most similar five-day sequence including the target day (n-lag = 4, Sect. 2.3.3.5, Fig. 2-10).

Monthly mean temperature fields are reconstructed separately using 22 stations providing monthly mean as predictor (Fig. 2-2). To allow seasonal shifts in monthly means, the analog pool is extended to the two straddling months (M3). This yields 150 possible analogs to reconstruct the monthly mean of a given month. The monthly mean T2M fields are interpolated in time to daily values using a sliding monthly mean with window length 2. The daily T2M anomalies reconstructed from the SLP predictor were then added onto the interpolated values from the monthly T2M reconstruction to complete the T2M reconstruction. This reconstruction thus includes the low frequency variations provided by the monthly station data and the high frequency variability provided by the daily SLP.

2.2.4 Testing the robustness of the analog-method

In order to assess the effect of modified settings of the AM on the reconstruction, several test cases are evaluated here. In a first test case, the idealized performance of the AM is evaluated within the surrogate climate of the RCAO simulation (RCAX in T2M case). Instead of using real station SLP as predictor, time series of daily SLP from model grid points in the vicinity of the real stations are used as ideal pseudo predictors for the reconstruction. The correlation between the reconstructed fields and the reference fields of RCAO can be taken as benchmark of the AM's optimal performance regarding temporal correlation. The ideal skill is compared with the correlation based on real SLP predictors in Sect. 2.3.1.

In order to estimate the robustness of the AM, a second test evaluates the sensitivity of the AM on the size of the analog-pool covering different periods. Whereas the final reconstruction of HiResAFF is obtained using the full 50 yr of analogs (case A), in this test the pool is divided in two parts, with the first 25 yr (B1 = 1958–1982) followed by a test using the second 25 yr (B2 = 1983–2007) as analog-pool, respectively. In case C, the pool is divided in 10-yr segments yielding five tests (C1 = 1958–1967, C2 = 1968–1977, C3 = 1978–1987, C4 = 1988–1997, C5 = 1998–2007). In all cases, the AM is used to reconstruct the 50yr covered by the reference data using standard settings. The retrieved correlations for the different test cases are shown in Sect. 2.3.2.1.

In a third test, the robustness of the AM is further evaluated by estimating the density of suitable analogs for the reconstruction of HiResAFF. In this test, we replace the best analog by the next neighbouring analog, then by the third, etc. till the n-th best neighbour. For every

next neighbour we calculate the correlation between the reconstruction based on increasingly poorer analogs and the reference data (Sect. 2.3.2.2). The mean field correlation of the different tests are depicted in Fig. 2-4. The slope of the decay in the correlations as a function of the rank in similarity of the chosen analog gives an estimation about the density of suitable analogs. If the slope is relatively small, the confidence in finding appropriate analogs is higher because the neighbouring analogs are quite similar to the best analog. A steep slope, in contrast, indicates a lower density of similar analogs .The AM is then not able to easily find analogs that are similarly good as the best one.

Finally, a fourth test evaluates the dependency of the reconstruction skill of the AM on changes of the number and spatial distribution of predictors. Test cases in which the number of stations is artificially diminished are defined and compared to the sixth case, in which all available stations are used in order to estimate the increased uncertainty using less stations at different locations (Fig. 2-5, Sect. 2.3.2.3).

2.2.5 Validation

The evaluation of the test cases and the validation of the reconstruction of HiResAFF are done by comparing the reconstructed fields with those of the RCAO simulation (RCAX in T2M case). In the temperature case, the reconstructed fields combine the information of two different models: daily anomalies of T2M from the RCAO model and the monthly mean T2M from RCA (Sect. 2.2.3.4). To avoid introducing an artificial bias in the validation, the reconstructed temperature was benchmarked against temperature data from both models, RCAO and RCA, combined in the same way as in the reconstructions. This reference field is denoted hereafter as RCAX. The rationale to validate the AM using the data from regional model simulations is to sideline the possible deficiencies of the AM itself. Using other independent data to benchmark the AM would automatically include a contribution of the RCAO model bias, which in principle is an independent source of error not related to the AM (Sect. 2.4.4). Using a leave-one-out approach, skipping always the actual year, the comparison between reconstructions and the reference dataset does not include an artificial skill. The validation is applied for the period 1958–2007 covered by the simulation with exception of T2M, where only the period 1961– 2007 is available for the reference fields.

Pearson correlation on daily and monthly scale is used to evaluate temporal covariance between reconstructions and the reference fields. Non-parametric Spearman rank correlation is additionally used in the daily precipitation case and for wind speed due to their non-normal distributions. Significance levels are estimated from 2-sided t-tests for $p<0.05$ and by $\pm Z_{(1+p)/2} \cdot \sqrt{N-1}$ with $Z_{(1+p)/2}$ being the (1+p)/2-quantile of the standard distribution in the rank correlation case (von Storch and Zwiers, 1999).

For every variable, the ratio of the variance ϕ = var(REC)/var(REF) of the reconstruction and the reference fields from RCAO is used for the evaluation of the reconstructed variance on daily and monthly scale. A 2-sided F-test is used for the estimation of significant deviations with $p<0.05$. For non-normally distributed variables of precipitation and for wind speed, the significance levels are derived by the bootstrap method (cf. Efron, 1982), including 1000 iterations for each N = 1500 samples. More specifically, a moving blocks bootstrap is used to consider the effect of the serial correlations in the daily data for precipitation (block length = 2) and wind speed (block length = 3) (cf. Liu and Singh, 1992; Ebisuzaki, 1997). The block length is estimated here based on the lag at which the autocorrelation of the daily variables becomes <0.2.

Significance levels for mean difference of the reconstruction minus reference fields (bias) is estimated from a 2-sided t-test with $p<0.05$ and from bootstrapping for variables with non-normal distribution. In order to test the deviation of higher quantiles, significance levels are estimated using "m out of n" bootstrapping, with $m = 2/3n$ to attribute the discontinuity of the distribution at higher quantiles (Cheung and Lee, 2005). For the high percentiles, a block length of one is used.

When conducting the same statistical test over a large set of theoretically independent grid-cells, a proportion of p grid-cells will yield false rejections of the null-hypothesis at the p significance level even when the null-hypothesis is correct. To test if the null-hypothesis as a whole (e.g. of no correlation between the target and the reconstructed multivariate fields) can be rejected, the field significance should be assessed (Livezey and Chen, 1983). This is performed by counting the number of local tests that surpass the local p significance level and comparing with the p-quantile in the distribution of the numbers of grid-cells that would surpass the local p level of significance under the null-hypothesis. In practice this can be accomplished by bootstrapping. When assessing the field significance of correlation between reconstructions and target field, 1000 bootstrap samples of the reconstructions are generated by re-sampling (and thus destroying the possibly existing time correlation), and calculating the correlation with the target field. The number of grid cells N_k (k=1,1000) that surpass the local p-level of significance are counted. The 95^{th} quantile of the distribution of N sets is the 95^{th} significance level. If the number N_{real} determined from the actual reconstructions is higher than the 95^{th} quantile, field significance can be claimed.

To test the bias and the ratio of variances, the distribution of N under the null-hypothesis is constructed by subsampling from the reconstructions 1000 bootstrap samples and counting the number of grid-cells in which the difference (or ratio of variances) surpasses the local significance level. When the variable can be assumed to be normally distributed, the local significance level is calculated from the corresponding theoretical expressions (Fischer Z, t-test, F-test). For variables that are potentially not normally distributed, like daily precipitation, the local p-significance level is determined by standard bootstrapping of the local series of reconstructions and the target variables as described before. On daily scale, serial correlation is taken into account by adjusting the block length for the bootstrapping for each variable, namely 6 for SLP, 3 for wind speed, rel. humidity and cloud cover, 2 for precipitation and 10 for T2M in January and 7 for T2M in July, depending on their autocorrelation, respectively. The test results are presented in Tables 2-1–2-3, respectively.

2.3 Results

2.3.1 Performance of the analog-method in the surrogate climate

For the evaluation of different reconstruction methods, state-of-the-art climate simulations provide a very useful surrogate climate of physically consistent atmospheric fields. Using model grid points as pseudo-predictors, the optimum reconstruction skill of a method can be estimated by comparing the reconstruction with the "truth" known from the model simulation. As the model presents only an idealized and simplified estimation of the real world, the idealized testing of the method might yield somewhat optimistic skills, e.g. because of a lower spatial variability in the model compared to observations. Nevertheless, this idealized testing of the method provides a good benchmark for the AMs potential skill based on the chosen settings.

In the AM case used for upscaling, three sources of uncertainties can be considered. The first one relates to the noise contained in the predictor data. This noise includes error measurements or local station variability that is not related to the predictand. A pre-filtering of the predictor dataset by empirical orthogonal functions (EOF) can be applied to separate the signal from the noise in the predictor. In the 23 station case providing daily SLP, this approach slightly decreased the reconstruction skill regarding correlations (not shown). Even if the predictor data would have been perfectly measured, a second source of error stems from solving Eq. (1) and finding only the most similar analog present in the archive, but not a completely equal pattern of predictors. As shown in Sect. 3.2.1 below, little improvement is achieved when increasing the analog-pool. Also, the density of suitable analogs (Sect. 3.2.2) indicates that the availability of analogs is saturated for the reference dataset used in this validation.

The third aspect relates to the linkage of real predictor data to the simulated predictor data taken from model simulations. While the relationship between the SLP predictor from grid points and corresponding predictand fields in the model world is consistent with the model physics, this cannot be expected when real station SLP is linked to the model fields (Fig. 3).

In order to estimate the theoretical optimal performance of the AM, the surrogate approach using grid point SLP from the model (grid points in the vicinity of real stations used in HiResAFF) is compared with the reconstruction obtained using real station SLP (Fig. 3). The correlation in the surrogate climate approach (case Ref) yields clearly higher correlations compared to case A using real station data. The difference of the explained variance between both cases r 2(Ref) - r 2(A) for SLP, wind speed and precipitation on daily and monthly scale is up to 10 % in January. In July, the difference is 25 % (17 %) for daily (monthly) SLP, around 12 % (18 %) for wind speed and 7 % (13 %) for precipitation, respectively. The large loss in the explained variance when linking real station data to model fields need to be kept in mind in the following evaluation, as it a priori lowers the reconstruction skill of the AM in this case, dependent on the used model (see Sect. 4.1).

2.3.2 Robustness of the analog-method

2.3.2.1 Dependency on the size and period of the analog-pool

For the application of the analog-method, an important question is how many analogs are needed for a successful reconstruction for a given domain (cf. van den Dool, 1994). To answer this question, sensitivity tests have been conducted in which the size of the archive has been varied. Changes in the correlations of reconstructions over the whole reconstruction period are used as one objective measure of the skill of the reconstructions.

Figure 2-3 shows the comparison of the reconstruction skill when eight different analog pools are used. Obviously, correlations do not considerably change when the analog-pool consists of only 10 yr (cases C) compared to the full size of 50 yr (HiResAFF, case A). Only the correlation of monthly means/sums tend to be slightly higher for case A than when using 10 yr as in cases C. Basically, the same results are achieved from a cross calibration and validation of 25 yr vs. 25 yr and 10 yr vs. 10 yr of different sub periods (not shown).

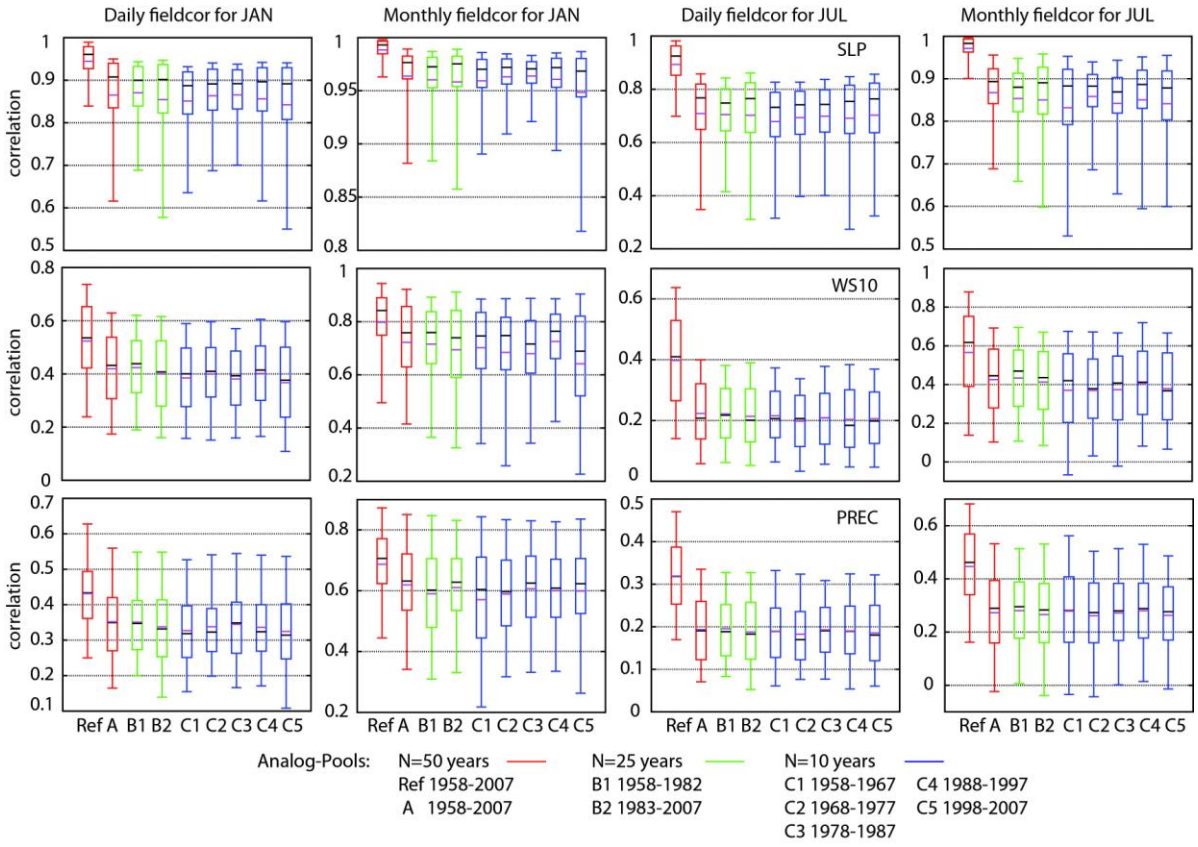

Fig. 2-3. Whisker-Box-Plots showing the field correlations of different test cases based on reconstructions from analog-pools of different size and periods. The box indicates the range of local correlations between the first and third quartile representing 50 % of the local correlations around the median (black horizontal line). The pink line represents the mean of field correlation. The whiskers indicate the spread of the correlations containing 90 % around the median. The reference case (Ref) is based on using model SLP as pseudo-predictor. The same stations are used for HIRESAFF (case A) but with real SLP. B1 and B2 use only 25 yr, C1–C5 only 10 yr from different periods, respectively.

2.3.2.2 Density of suitable analogs

Independent from the size of the analog-pool, the availability of suitable analogs within a given pool is evaluated by searching the n-th best analog instead of just the best analog. In this test the archive size was always 4500 days. The decay in the mean field correlation as a function of the analog rank is shown in Fig. 4. Whereas the decrease in correlation is rather rapid when going from the 10th to 50th best analog, the slope becomes rather linear for higher ranks. As an example for reconstructed daily SLP, the explained variance decreases linearly with a rate of around 6 % per 100 neighbours in January and 3 % for July for neighbours > 100 to the best analog. For the first ten neighbours, the slope is larger with already a decrease of 6 % in January and 7 % for July for 10 neighbours, respectively.

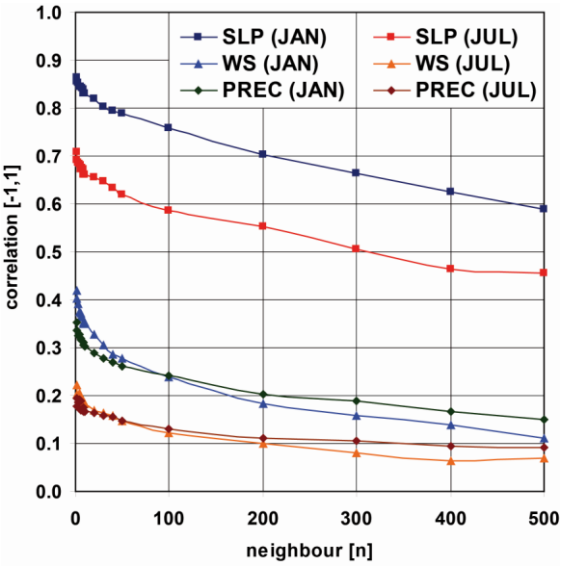

Fig. 2-4. Density of suitable analogs for HiResAFF estimated as the decay in daily mean field correlation as a function of next neighbours replacing the best analog chosen from around 4500 possible analogs. Displayed are the variables of SLP, wind speed (WS) and PREC for January and July.

2.3.2.3 Dependency on the number of predictors

In order to test the predictive skill when the number of predictors is reduced, six test cases are shown in Fig. 5. To avoid the effect of missing values contained in the station data when using a reduced number of stations, the tests are based on model grid points of SLP instead of station data. The used grid points for the different tests are shown in Fig. 5. The correlations of the reconstructions with the reference fields are shown for daily wind speeds for January and July. Only the reconstruction skill for daily wind speed is presented here as an example for a variable with a strong physical link to SLP but with a high spatial variability.

In Fig. 5c1, the results have been obtained with three predictors located over the central and southern Baltic Sea. The correlation of daily wind speed with the reference fields shows already high values of r>0.5 within the triangle spanned by the location of the three predictors for January and July. Adding a fourth grid point in the north (Bodø, 67°25' N, 14°25' E) in Fig. 5c2 largely extends the area with improved correlations with at least r>0.4 in January, whereas the improvement is low in July. Test cases c4 and c5 show an example where the whole field is reconstructed by using 5 grid points close to the boundaries in c4 and an additional grid point in the centre in c5. Test case c5 shows a large improvement of the median field correlation (r = 0.40) compared to c4 (r = 0.33) in January where only the predictors are all located at the boundaries of the domain. For July, the improvement in c5 is reflected in broader areas with correlations exceeding at least r>0.2 compared to c4. However, the very low (c5) to non-significant (c4) correlations at the eastern boundary in July can even persist at the locations of the grid points used as predictor.

While the former test cases were rather artificially constructed, test case c3 shows the reconstruction skill for six grid points representing the situation of the available real data in 1850. Consequently, low correlations can be expected on daily scale at the boundaries, with no significant skill in July for the northern and north-western boundaries. Finally, test case c6 represents the skill of the reconstruction in the surrogate climate when all 23 grid cells (surrogate of the 23 stations) are available (corresponding approx. to the period 1870–1990). It should be noted that using real SLP instead of grid point SLP from the model yields generally lower correlations but similar spatial patterns, as shown in Fig. 3 for the comparison of case Ref (23 pseudo predictors) with case A (23 real stations),

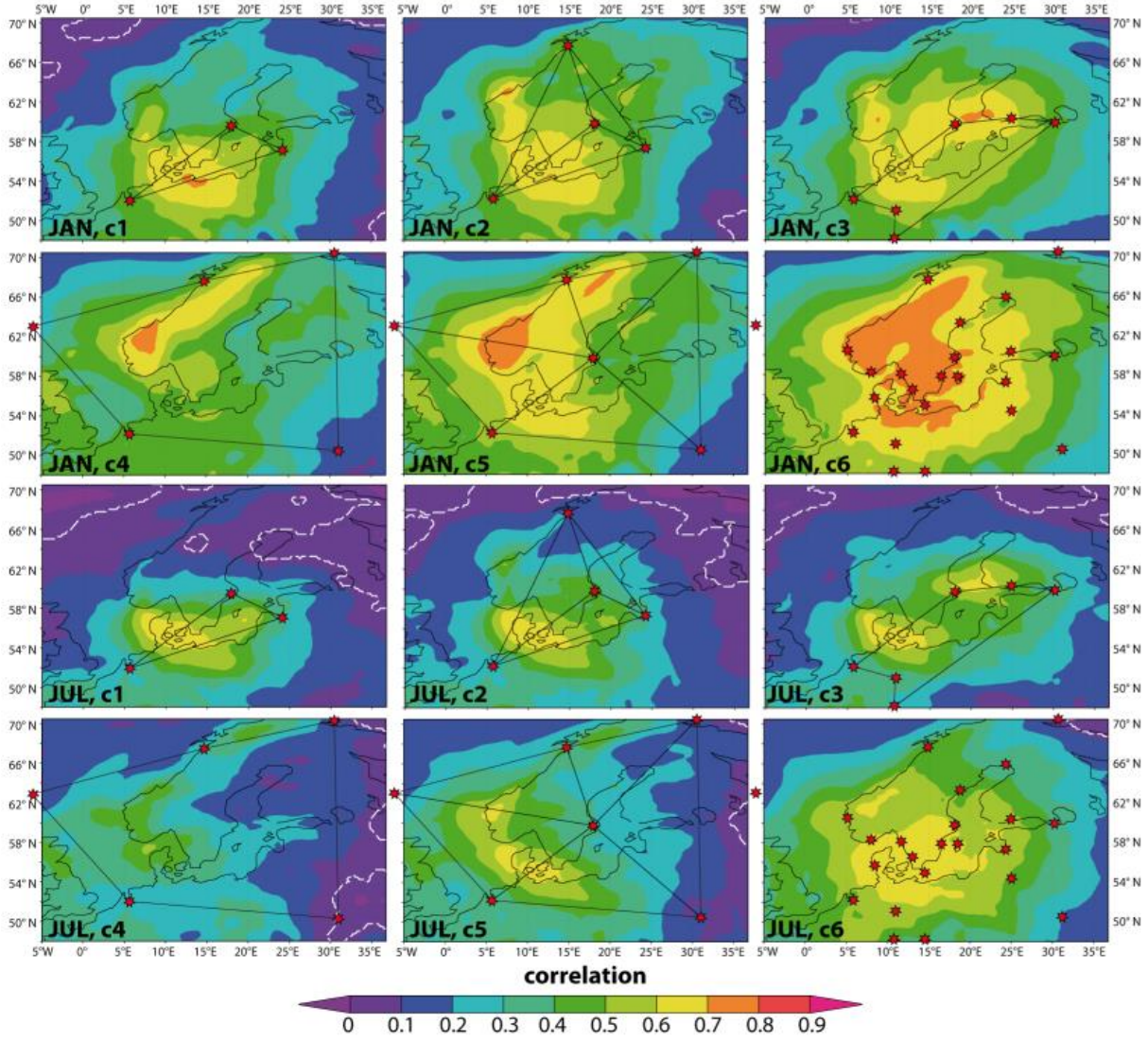

Fig. 2-5. Correlation of daily wind speeds between a surrogate reconstruction and reference fields of RCAO (RCAX for T2M) for January and July 1958–2007 dependent on the number and distribution of SLP predictors. White shaded lines indicate areas where h_0 of zero correlation cannot be rejected with $p<0.05$.

2.3.3 Validation of HiResAFF for the period 1958–2007

The reconstructed fields of HiResAFF using the standard settings described in Sects. 2.3.2 and 2.3.4 are validated with the reference fields from the RCAO simulation (RCAX for T2M) on daily and monthly scale. In the following only January and July are presented, as reconstruction skills are highest in winter and lowest in summer with other months in between.

2.3.3.1 Correlation

The temporal correlation between HiResAFF and the reference fields for different variables on daily and monthly scale for January and July are shown in Fig. 6. The mean over all local correlations of the field are given in Table 1 together with the amount of local tests h [%] showing significant correlations with $p<0.05$. All variables are showing significant field correlations at the 5 % confidence level on daily and monthly scale, respectively.

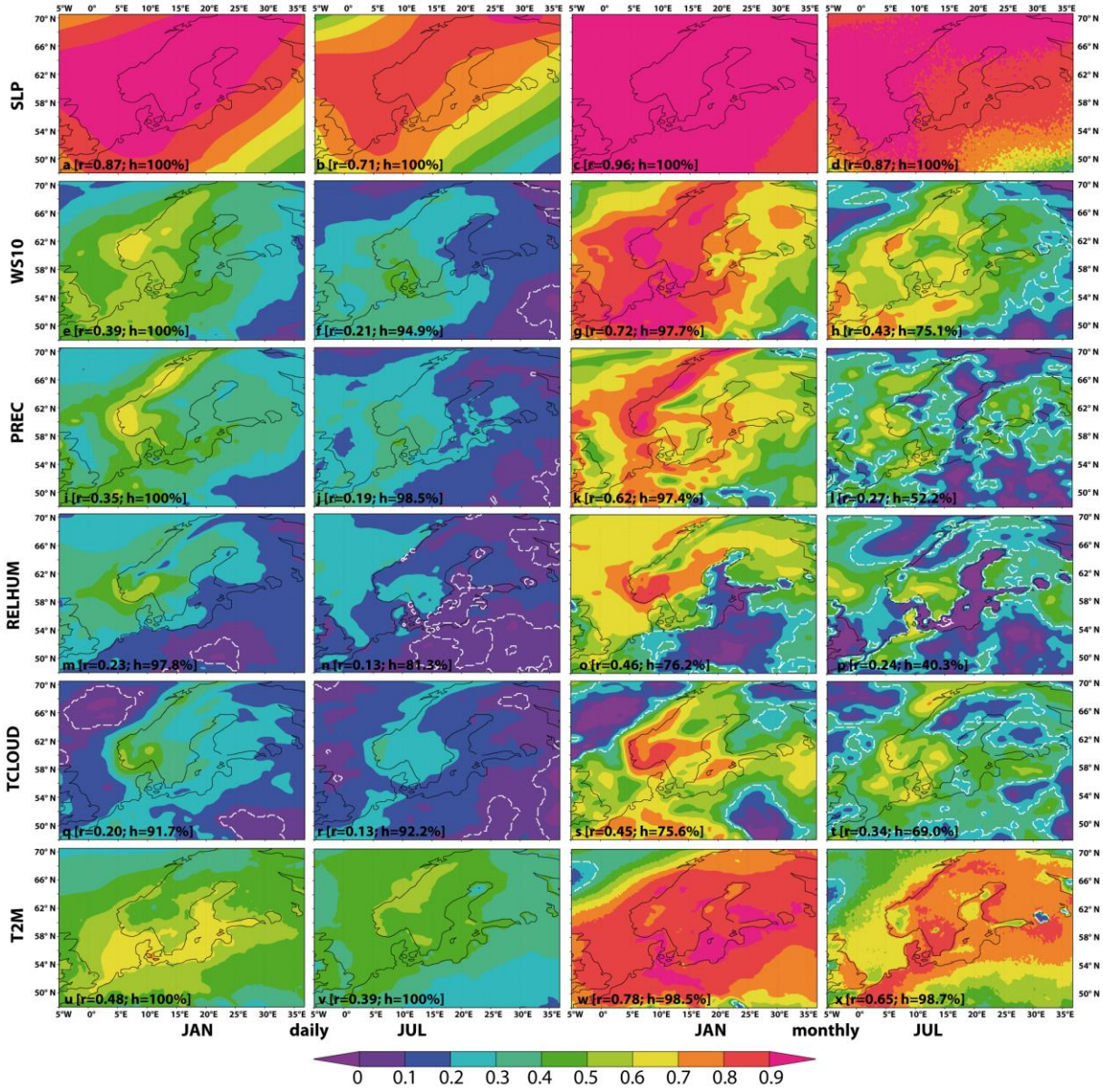

Fig. 2-6. Correlation maps on daily (left column) and monthly (right column) scale between HiResAFF and the reference fields of RCAO (RCAX for T2M) for January and July 1958–2007. White shaded lines indicate areas where h_0 of zero correlation cannot be rejected with $p<0.05$.

Very high correlations are generally achieved for SLP due to the strong physical link to the predictor and low spatial degrees of freedom of this large-scale variable. Lowest skills are evident over the south-eastern domain (Fig. 2-6a–d). Although less pronounced, the general feature of lowest correlation in this region is also found using model data as surrogate predictors in Sect. 2.2.4, even if an additional predictor is used in this region (not shown).

Daily correlations of wind speeds in January are all statistically significant at the 5 % level, with high values in the windward areas and lower values in the east and over the NW. In July, daily correlations show a similar dipole pattern, with high values in the west and low to non-significant correlations in the east. Correlations of monthly mean wind speeds show comparable spatial distributions in the field correlation with clearly higher skills on average, although in January the SE domain and in July also the NE-Atlantic and most parts of the eastern boundary show non-significant values.

Daily precipitation in January shows generally significant correlations with higher values (r>0.4) for windward coastal and mountain areas, with decreasing skill towards the eastern and SE domain. A similar spatial pattern for the correlation of monthly precipitation amounts in January is achieved with generally good correlations. Daily precipitation in July is reconstructed with higher correlation over the western and the central domain, but with low to non-significant skills in the eastern part. Monthly amounts of precipitation in July show higher correlations over the western and partly eastern domain but non-significant values for northern and south-eastern regions and the Baltic Sea.

Daily correlations of relative humidity in January are mostly significant with a dipole pattern of high values in the NW vs. low values in the SE. This is also the case for correlations of the monthly mean in January, with non-significant correlations in the SE domain and most parts of the Baltic Sea. Daily correlations of relative humidity in July are generally very low, with higher values over the western domain and low to non-significant values over the eastern domain. Monthly mean humidity in July shows higher correlation over land and the NE-Atlantic but low and partly negative correlations over the Baltic Sea, the SE domain and UK.

Reconstructed daily cloudiness in January shows non-significant correlations over the NE-Atlantic and the SE, with highest values over the windward coastal areas. A similar pattern exists also for the monthly mean cloudiness, with much higher correlations except for the NE-Atlantic and the SE. Daily correlations of cloudiness are very low in July, with slightly higher values for windward coastal areas. The correlations for the monthly mean cloudiness in July show higher values for the central domain with a general heterogeneous pattern with non-significant values, i.e. over the NE-Atlantic and N-Scandinavia and southern regions.

Table 2-1. Mean correlation between RCAO (RCAX for T2M) and HiResAFF on daily (left) and monthly (right) scale for January and July. Additionally shown is the amount of local tests h [%] showing significant correlations with p<0.05. Field significance with p<0.05 can be claimed for all cases according to bootstrapping test.

	JAN (daily)		JUL (daily)		JAN (monthly)		JUL (monthly)	
	cor(d)	h [%]	cor(d)	h [%]	cor(m)	h [%]	cor(m)	h [%]
SLP	0.87	100%	0.71	100%	0.96	100%	0.87	100%
WS10	0.39	100%	0.21	94.9%	0.72	97.7%	0.43	75.1%
PREC	0.35	100%	0.19	98.2%	0.62	97.4%	0.27	52.2%
RH	0.23	97.8%	0.13	81.3%	0.46	76.2%	0.24	40.3%
TCC	0.20	91.7%	0.13	92.2%	0.45	75.6%	0.34	69.0%
T2M	0.48	100%	0.39	100%	0.78	98.5%	0.65	98.7%

2.3.3.2 Variance

The ability of the AM to realistically reconstruct the high frequency daily to monthly variability is evaluated by calculating the ratio of variance (rv) between HiResAFF and the reference fields $rv = \phi = \sigma_{HiResAFF}/\sigma_{RCAO}$. Figure 2-7 shows the ratio of variance σ for the different variables for January and July on both time scales. The field average of ϕ and the number of local 2-sided tests h [%] for which the null-hypothesis of no significant deviation in variance have to be rejected at a significance level of p<0.05 are given in Table 2-2.

Daily variance of SLP tends to be slightly underestimated in the reconstruction, with significant underestimations at the eastern boundary in January and the central to western Baltic Sea in July. Variability on monthly scale in January and July shows a strong underestimation, i.e. over the SE domain.

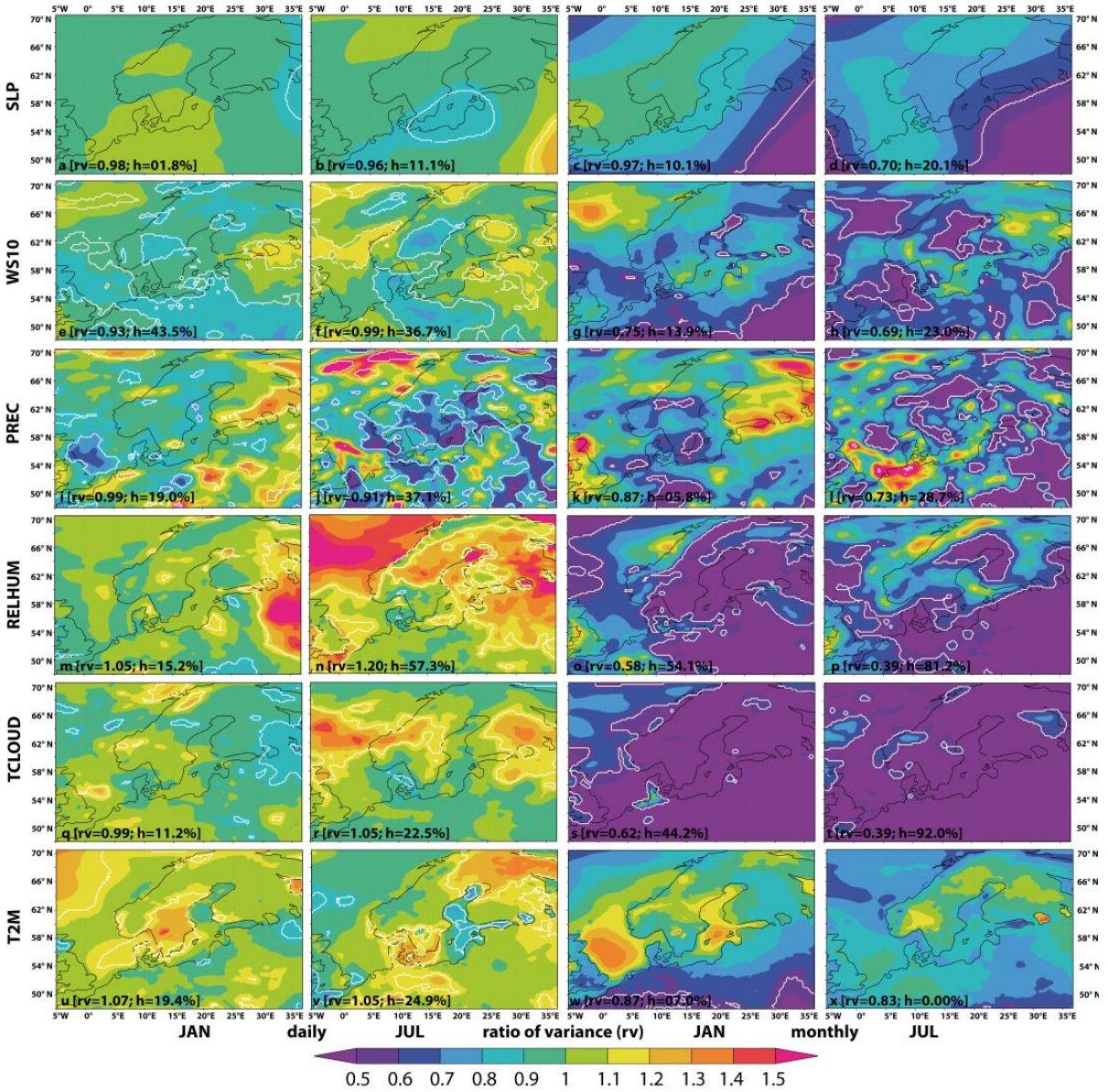

Fig. 2-7. Ratio of variance on daily (left column) and monthly (right column) scale between HiResAFF and the reference fields of RCAO (RCAX for T2M) for January and July 1958–2007. White shaded lines indicate areas where the reconstruction shows significant deviations.

The variance of daily wind speeds tends to be mostly underestimated in January, while mostly significant deviations in variance of both signs are reconstructed for July. On monthly scale, regions with too low reconstructed variance dominate in January at the southern and eastern boundary, over the North Sea and those parts of the Baltic Sea being usually covered by sea-ice. Realistic variances are reconstructed over most areas of the central domain, with slightly overestimated variance over the NE-Atlantic. For July, the reconstructed variance on monthly scale is underestimated with heterogeneous spatial distribution.

Variances of daily precipitation are on average realistically reconstructed for January, with significant underestimation i.e. over the North Sea, and overestimation over continental regions in the E and SE. For July, variance of daily precipitation shows regionally very heterogeneous under- and overestimations. Variances of the monthly precipitation amounts for January tend to be slightly underestimated in the reconstruction, with higher variance over the SW and NE and lower variance over the central and S domain. In July, variance is underestimated with spatially heterogeneous deviations.

Variance of daily humidity in January is, on average, realistically reconstructed with exception of significant overestimations in the E domain. For July, the daily variance is overestimated for large areas over the NE-Atlantic and Fennoscandia, with more realistic values in the central and southern domain. On monthly scale, variance in January is strongly underestimated (ϕ <<50%), in the central and E domain, with more realistic values only over the SW, North Sea and partly along the Norwegian coast. Monthly variance of humidity in July shows very strong underestimation (ϕ <<50%), i.e. over the SE of the domain.

Table 2-2. Mean ratio of variance ϕ between HiResAFF and RCAO (RCAX for T2M) on daily (left) and monthly (right) scale for January and July. Additionally shown is the amount of local tests h [%] showing significant deviations of ϕ with p<0.05. Values in italics show no field significance for the deviation in variance with p<0.05.

	JAN (daily)		JUL (daily)		JAN (monthly)		JUL (monthly)	
	ϕ (d)	h [%]	ϕ (d)	h [%]	ϕ (m)	h [%]	ϕ (m)	h [%]
SLP	*0.98*	*1.8%*	*0.96*	*11.1%*	0.97	10.1%	0.70	20.1%
WS10	0.93	43.5%	0.99	36.7%	0.75	13.9%	0.69	23.0%
PREC	*0.99*	*19.0%*	0.91	37.1%	*0.87*	*5.8%*	0.73	28.7%
RH	*1.05*	*15.2%*	1.20	57.3%	0.58	54.1%	0.39	81.2%
TCC	*0.99*	*11.2%*	1.05	22.5%	0.62	44.2%	0.39	92.0%
T2M	*1.07*	*19.4%*	1.05	24.9%	*0.87*	*7.0%*	*0.83*	*0.0%*

Daily variance of cloudiness is reconstructed realistically, on average, with a slight tendency to underestimation in the E-NE domain in January. In contrast, the regions in the ENE show overestimated variance in addition to large parts of the NE-Atlantic in July. On monthly scale, variance is clearly underestimated, with exception of the northern and SW domain in January. For July, variance is strongly underestimated (ϕ <<50%) for all regions.

The daily variance of the T2M reconstruction of January and July is realistically reconstructed with regional deviations of both signs. In July, the daily variability is underestimated i.e. over most parts of the Baltic Sea and the NE-Atlantic, while overestimated on land. Variances on monthly scale in January and July are underestimated but with mostly non-significant deviations. However, for January the southern boundary shows a significant underestimation in variance while being too high over the North and Baltic Sea.

2.3.3.3 Reconstruction bias

The bias in mean (monthly sum for precipitation) $\Delta\overline{m} = \overline{m}_{HiResAFF} - \overline{m}_{RCAO}$ of the reconstruction is shown in Fig. 2-8. The average bias of the field 1m and the number of local tests h [%] showing significant deviations with p<0.05 are summarized in Table 2-3. Reconstructed SLP fields show no significant difference in mean for January. The east–west dipole indicates up to 0.4 hPa too high mean SLP over the Norwegian Sea and slightly too low values in the eastern part. In July, however, SLP in the SW domain is significantly too low (down to -1.7 hPa) while values in central NE domain are significantly too high.

With exception of the NE domain, wind speeds tend to be in general significantly underestimated, i.e. over oceanic regions. Wind speeds are generally underestimated in the central domain and mostly pronounced over the North Sea, while the NW and SE domain shows significantly higher wind speeds.

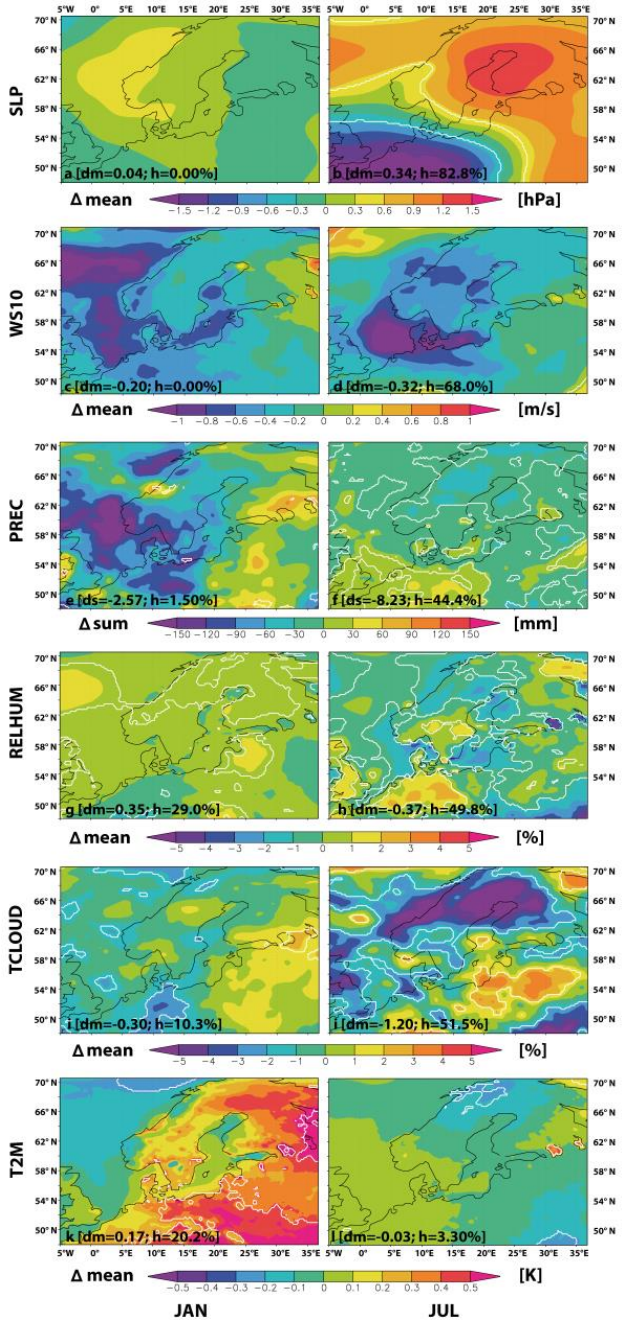

Fig. 2-8. Mean bias of HiResAFF minus the reference fields of RCAO (RCAX for T2M) for January (left column) and July (right column) 1958–2007. White shaded lines indicate areas where the re-construction shows significant bias in mean with p<0.05.

Reconstructed precipitation amounts show mostly non-significant deviations in January, with lower values over the seas and too high precipitation amounts over continental areas towards the E. In July, precipitation amounts are underestimated, i.e. over Fennoscandia, and overestimated over central Europe.

Relative humidity shows a tendency to overestimation in January. In July, deviations show a spatially heterogeneous picture dominated by regions with significant underestimation, i.e. over the seas. Mean total cloud cover is underestimated in January over most areas, with exception of the E-SE domain showing overestimation of 1–2 %. Deviations in mean in July show a heterogeneous picture with significant deviations in both directions and an overall tendency to underestimate cloudiness.

Mean T2M in January shows a warm bias, i.e. over land most pronounced in the eastern and southern domain, while T2M over seas show only small deviations. T2M in July shows generally small non-significant deviations with a tendency to a small cold bias.

In addition to the bias in the mean, we calculated also the deviation of higher percentiles of the reconstruction minus the reference fields for daily wind speed and precipitation for the 50-yr period. Using the "m out of n" bootstrap (Sect. 2.2.5) to estimate significant deviations of higher percentiles, we find no significant deviation, with $p<0.05$ for the 90th, 95th or 99th percentile (not shown), while significant deviations partly occur around the mean value (Fig. 2-8).

The realistic reconstruction of extremes can be explained by the AM's ability to reproduce the correct frequency distributions of different variables (Zorita and von Storch, 1999; Fernández and Sáenz, 2003), which is demonstrated below.

Table 2-3. Difference in mean (bias) between HiResAFF and RCAO (RCAX for T2M) for January (left) and July (right). Additionally shown is the amount of local tests h [%] showing significant bias with p<0.05. Globally non-significant bias with p<0.05 is indicated by italics.

	JAN		JUL	
	Δmean	h [%]	Δmean	h [%]
SLP	*0.04 hPa*	*0.0%*	0.34 hPa	82.8%
WS10	*-0.20 m/s*	*0.0%*	-0.32 m/s	68.0%
PREC	*-2.57 mm*	*1.5%*	-8.23 mm	44.4%
RH	0.35 %	29.0%	-0.37 %	49.8%
TCC	-0.30 %	10.3%	-1.20 %	51.5%
T2M	0.17 K	20.2%	-0.03 K	3.3%

2.3.3.4 Frequency distributions

The ability of the AM to reproduce the frequency distribution of the different meteorological variables of HiResAFF is shown in Fig. 2-9 for January and July, respectively. The "true" reference distributions of RCAO (RCAX for T2M) are shown as shaded lines compared to the distributions reconstructed at the same location in HiResAFF (solid lines). In all cases, the general distribution types are clearly reconstructed using daily SLP as predictor, including the upper and lower tails and extremes.

In the SLP case, examples of frequency distributions are shown for grid points showing the latitudinal changes in the distribution between de Bilt (52° N, 5.25° E) vs. Haparanda (65.5° N, 24° E). The reconstruction clearly reproduces the different climate regimes regarding circulation, with prevailing westerly flow and high occurrence of lows in high latitudes visible in the broader distribution and the shift towards lower pressure (Haparanda), compared to de Bilt showing a more narrow distribution shifted towards higher pressure, i.e. in July. For the other variables, distributions at two grid points are shown as examples focusing on meridional changes between Bergen (60.25° N, 5.25° E) – representing maritime-advective conditions – and St. Petersburg (60° N, 30.25° E) – representing more continental conditions. As indicated by the embedded scatter plots in Fig. 2-9 for the different variables, a linear regression of the frequency distributions of the reconstruction with those of the RCAO yields slope parameters very close to 1 with explained variances $r^2 > 0.95$, with exception of January T2M in St. Petersburg ($r^2 = 0.91$).

While the reconstructed frequency distributions of wind speeds do not show systematic deviations, the frequency of wind directions slightly differs around the main wind directions. In the example of Bergen, the wind direction in January tends towards more SSE and SSW direction compared to the reference fields, while SE directions are slightly underestimated. Also for the St. Petersburg case, wind directions from WSW in July are overestimated in the reconstruction compared to the reference fields. However, it should be noted that the large bins of 22.5° in the wind rose make the frequency counts sensitive to small directional changes between neighbouring bins, i.e. around the main wind directions.

The discontinuous distribution of relative humidity partly shows deviations for high values, e.g. in the case of Bergen for July. The extremely high frequency of very high total cloud cover in January for St. Petersburg is fully reproduced in the reconstruction. The general distribution of daily precipitation is also reconstructed well. Note that the natural logarithm of the total frequency N is used here to highlight deviations of high to extreme precipitation events with low frequencies at the upper tail of the distribution. Due to the large bin intervals of 5 mm and the logarithmic scale for N, deviations between neighbouring bins appear larger,

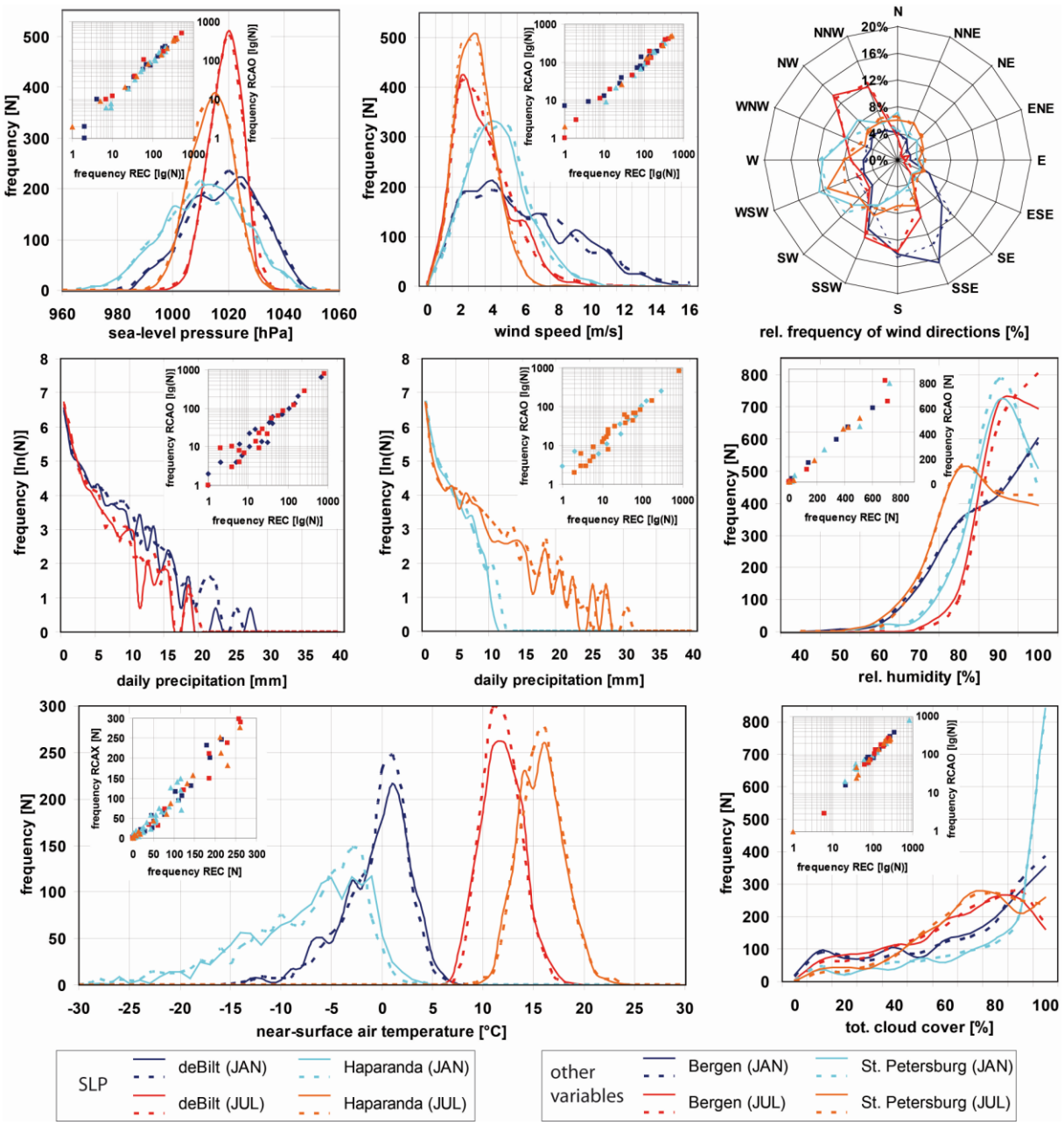

Fig. 2-9. Comparison of the frequency distributions between reconstruction (HiResAFF, solid lines) and reference fields from RCAO (RCAX for T2M) (shaded lines) for different variables for January and July. Grid points are chosen in the vicinity of de Bilt and Haparanda in the SLP case to highlight latitudinal changes. For other variables, Bergen and Saint Petersburg are chosen to depict differences between more maritime-advective (Bergen) and continental(Saint Petersburg) climate regimes. Embedded are the scatter plots of the regression between the reconstructed and simulated distributions. Note different usage of scaling (log, and ln).

while the general distribution does not deviate considerably. However, mismatches in the magnitude of strong precipitation events should be expected due to the locally very heterogeneous occurrence of strong rain events.

The frequency distributions for daily temperature based on the combined approach of multivariate predictors (Sect. 2.2.3.4) are in good agreement with the reference fields for the given examples. Note that small bins of 2 K are used for the calculation of the frequencies to highlight deviations around the mean value. The increasing warm bias of HiResAFF towards

the E (Fig. 2-8) is also visible for the T2M distribution of St. Petersburg in January. While the lower tail of the distribution of extremely cold to cold (T<-5° C) temperatures does not deviate considerably, the frequency of temperatures between -5° C and 0° C are clearly underestimated, while the right tail of warmer temperatures is overestimated, leading to the warm bias. As indicated also for the other T2M distributions including July in Fig. 2-9, largest deviations occur around the mean value leading to a broader distribution of the reconstruction compared to the reference fields.

2.3.3.5 Auto-correlation

Figure 2-10a shows the reconstructed auto-correlation of different variables compared to the reference fields for January and July. As an example, a grid point in the centre of the domain in the vicinity to Stockholm (59.25° N, 18° E) is chosen although other locations would show little difference. In the SLP case, the serial correlation is almost realistically reconstructed with only a slight underestimation. In the daily wind speeds case, serial correlation is at least partly reconstructed but clearly lower than in the RCAO simulation. The already very low persistence in daily precipitation is reconstructed in January but not in July. For relative humidity and total cloud cover, the AM fails to reconstruct the considerable persistence in the reference simulation.

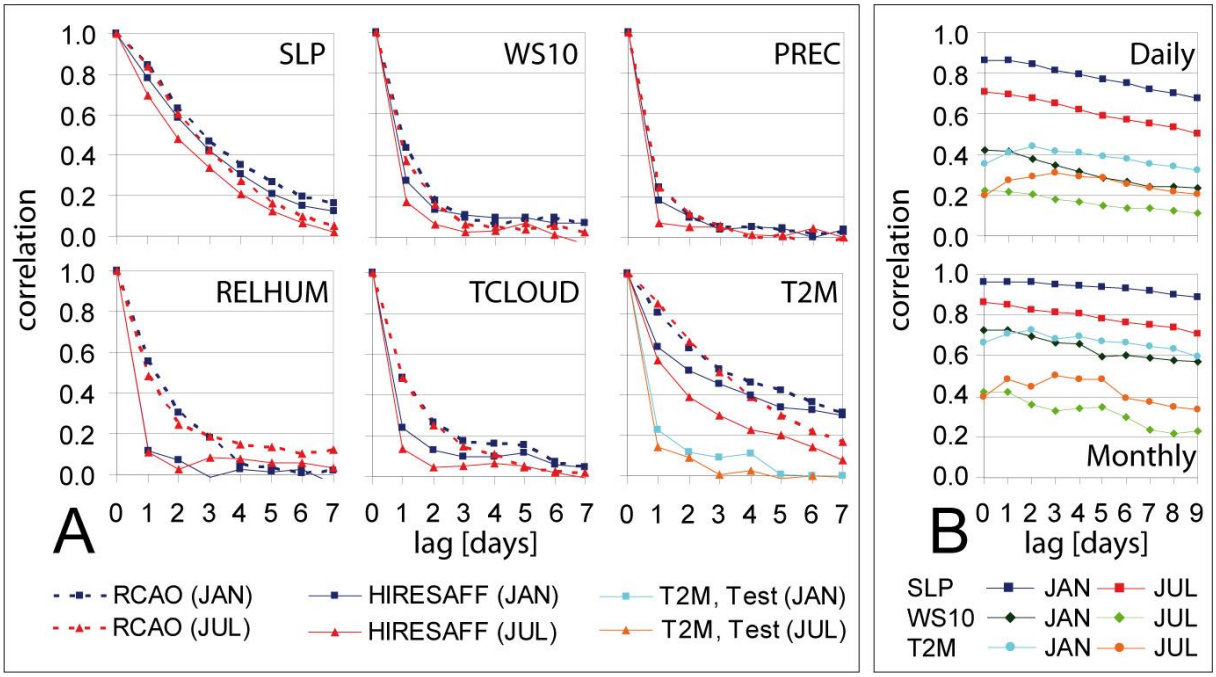

Fig. 2-10. Effect of implementing persistence in the analog-reconstruction. (A) Reconstructed daily auto-correlations of HiResAFF (solid lines) dependent on the used n-lag in the AM with RCAO and RCAX (shaded lines) for January and July at a grid point close to Stockholm. The test cases (light blue, orange) for T2M show the daily serial-correlation if T2M is reconstructed using the standard settings from Sect. 2.2.3.2. (B) Changes of daily (top) and monthly (bottom) mean field correlation between reconstruction and RCAO (RCAX) for different variables for January und July dependent on the used n-lag for the AM.

For daily T2M, two reconstructions are compared in Fig. 2-10a based on different settings used for the AM. In light blue and orange, the standard-setting T2M reconstruction is shown (Sect. 2.2.3.2) without implementation of persistence in the AM. In this case, the high serial-correlation of the SLP predictor does not carry over to high persistence in T2M. Hence, the AM is not able to reconstruct the important memory in daily T2M.

For this reason, the alternative temperature reconstruction of HiResAFF (Sect. 2.2.3.4) aims to replicate the observed persistence of the predictand by choosing the most similar succession in the previous n-lag = 4 days instead of only the best analog of the target day. Although it turns out that this approach still underestimates the persistence, the reconstructed autocorrelation shows a very clear improvement. Using n-lag > 4 further improves the daily persistence towards the simulated values (not shown). However, with increasing value of n-lag it also becomes increasingly difficult to find different analogs for two successions that differ only in one or two days. The result is that the method tends to identify the same analog for consecutive days, which is unrealistic. There is also a price to be paid for improving the time persistence in the reconstructions, since the selected analog sequence of days will not in general contain the best analog for the target day. As a consequence, the mean field correlation of the reconstruction with the reference field decreases with increasing n-lag value used in the AM reconstruction (Fig. 2-10b). The choice of the value of n-lag thus depends on a trade-off between achieving a good daily persistence and a smaller reconstruction error.

In the setting just described, all days in the sequences leading to the target day are weighted equally in the search for an analogue sequence. A compromise between the standard setting and the one just described is to weight the days in the sequence unequally, with diminishing weights applied to days farther apart from the target day. Here, a weighting scheme that is proportional to the observed serial correlation in the predictand has been applied. An example is the model grid point close to Stockholm. The autocorrelation over four days, normalized to yield a sum of 1, yields 0.45, 0.27, 0.17 and 0.11, respectively. A reconstruction with weighted n-lag = 4 yields an autocorrelation of the reconstruction of 0.63, 0.31 and 0 for the example of a grid point close to Stockholm in January. Although the autocorrelation strongly improves for lag 1 day, it strongly decays to 0.3 for lag 2 days and disappears for a lag of 3 days. In contrast, when using equal weights for all n days in the sequence, the autocorrelation improves with a much slower decay. For the example in Fig. 10a for January, equally weighted n-lag = 4 yields an autocorrelation for T2M of 0.70, 0.48, 0.34 and 0.21.

2.4 Discussion

2.4.1 Analog-upscaling in the surrogate climate vs. observations

The comparison of the optimal performance of the AM in the surrogate climate of RCAO (case Ref.) with the reconstruction (case A) using real station data as predictor shows a clear loss in the explained variance when linking real data to model fields (Fig. 2-3). In Sect. 2.3.1 we separated three different sources of errors which might affect our reconstruction. As a first aspect, the data quality of the station readings does not seem to explain the large discrepancies to the surrogate climate approach. When the station data are pre-filtered by an EOF analysis, truncating the data by retaining only the leading EOFs, the reconstruction skills do not change much. Also, the second aspect of having not enough suitable analogs seems to be not relevant given that the reconstruction skills obtained with much smaller archives are also very similar (Fig. 3 and Sect. 3.2.1). The reason for the loss in the explained variance of the reconstruction compared to the optimal skill of the AM in the surrogate climate lies, therefore, in the third aspect of linking observations to model fields.

Since RCAO and RCA3 are driven only at their lateral and lower domain boundary by ERA40 and SST (in the case of RCA3), the model develops its own solution in the interior domain which may lead to differences compared to observations. The deviations will be generally larger in summer than in winter due to the reduced boundary forcing of the large-scale in the summer season (Déqué et al., 2007). Consequently, the discrepancy is larger in

July than in January (Fig. 2-3) when temperatures at mid and high latitudes are known to be less connected to the large-scale atmospheric circulation. Precipitation and cloudiness in summer are also strongly determined by small scale processes related to convection. As RCMs are designed to parameterize those processes, the simulation has more degrees of freedom on the local scale, possibly leading to deviations from observations in the interior of the model domain. In addition, one might speculate that these processes cannot be fully captured by the predictor field with a density of only 23 stations. However, as indicated by the surrogate approach using SLP data from 23 model grid points in the vicinity to the real stations as predictors, the skill of the AM is comparably high.

One possibility to reduce the gap between the model simulation and observations – and at the same time improve the correlation of the reconstruction, i.e. on daily scale – is the application of spectral nudging when numerically downscaling reanalysis data (e.g. von Storch et al., 2000, Yoshimura and Kanamitsu, 2008). This approach has been shown to bring the model closer to observations also in the interior of the model domain. However, like in the case of the used RCAO and RCA3 models, many regional models are not using spectral nudging so far. Our analysis can therefore be considered as a realistic application of the AM with the potential of considerable improvement through usage of spectral nudging for the regional climate simulation.

2.4.2 Robustness of the analog-method

The results in Sect. 2.3.2.1 and Fig. 2-3 show that the AM, and hence the reconstruction of HiResAFF, is very robust concerning the size of the analog-pool. The relatively small difference in the level of correlations achieved for a reconstruction when selecting the analog from 50 or only 10 yr demonstrates that the amount of analogs is more than sufficient, at least for the period 1958–2007. Similarly, the results suggest that further expanding the analog-pool will not considerably improve the reconstruction in this case. The similar level of correlations when using different sub-periods in Fig. 3 demonstrates that the reconstruction skill of the AM is independent from the choice of the calibration and validation period. The AM will remain stationary also on longer time scales as long as the climate (in absolute values or spatial patterns) does not evolve outside the range of observations in the analog-pool.

Testing the density of suitable analogs contained in the analog-pool (Sect. 2.3.2.2, Fig. 2-4) additionally confirms that the availability of suitable analogs is high for HiResAFF. Omitting always the best analog, which is already a drastic artificial deterioration when being used for every analog, would still lead to a relatively good level of correlation. In the same time, the steep slope in the correlations obtained with the first 10 to 50 neighbours indicates that the AM will to some extent depend on the analog-pool. This slope gives only a relative measure about the density of the analogs dependent on the current dataset. Comparing the slope with those achieved from a reconstruction of another region could be used to estimate whether different regions need larger analog-pools than other regions (e.g. dependent on the large-scale flow or topography etc.). The density might, therefore, be not only dependent on the size of the analog-pool but also on the size of the domain, the complexity and hence the geographical region.

The third test deals with the question of which effect the number of used predictors and their geographical distribution has on the reconstruction skill, namely correlation (Sect. 2.3.2.3, Fig. 2-5). For the presented example of daily wind speed in January und July, the results indicate that in winter a relatively small number of predictors of three to six yields already

promising skills due to the dominating large-scale forcing. While even regions not covered by the predictor show significant correlations in January, meso- to local-scale variations in summer yield considerable lower skills with mostly non-significant correlations in remote parts of the domain if only a reduced number of predictors is used. As the analog-upscaling always involves atmospheric fields from a regional climate model, testing of different numbers and locations of predictors provides a very helpful meta-test to find a suitable size and location of a domain. In addition, these test cases allow the estimation of uncertainties of a reconstruction related to, e.g. a decreased number of predictors such as missing values or less data back in time. As in the example c3 in Fig. 2-5 representing data availability in 1850, relatively high correlations can still be expected over the Baltic Sea region while little skill can be expected for remote regions, i.e. in summer.

2.4.3 Validation of HiResAFF

2.4.3.1 Correlation

The reconstruction skill of the AM regarding correlations on daily scale in January and July clearly shows a dependency from the westerly flow for all variables with exception of cloudiness. This can be explained by using SLP as predictor. Hence, correlations show a dipole pattern with higher values towards the W and lower values towards the E and SE. It should be noted that the higher correlations in W are achieved with a low station density in the western domain (Fig. 2-2) while the high station density in the central domain does not considerably improve the skill towards the eastern domain. Cloudiness is in contrast more dependent on the area covered by a higher density of stations. The temperature reconstruction yields relatively good correlation skills on daily scale although daily anomalies are only predicted by daily SLP with implemented persistence of 4 days (Sect. 2.2.3.4).

On monthly scale, the E–W dipole pattern displayed by the correlations is also visible for SLP, wind speed, precipitation and partly relative humidity in January. In July, small-scale convective processes lead to spatially more heterogeneous skills for precipitation and humidity with no skill over the Baltic Sea for humidity. Reconstructed monthly mean T2M shows very high (January) to high (July) correlations over land with low skill over the NE-Atlantic in January and additionally the North Sea in July. This can be explained by the chosen predictor data of monthly mean T2m that reflects the temperature on land apart from rather slow and therefore differing changes in sea-surface temperatures of the North Atlantic or the North Sea. The correlations obtained for monthly cloudiness are satisfactory given that no suitable predictor is available for the reconstruction.

2.4.3.2 Variance

Based on the results in Sect. 2.3.3.2, it can be concluded that the AM yields on average realistic values of reconstructed high-frequency variability on daily scale. Hence, the advantage of the AM of no loss in variance in the reconstruction for downscaling precipitation (Zorita and von Storch, 1999; Fernández and Sáenz, 2003) is valid also for upscaling on daily scale for different variables. However, on monthly scale, the variance is on average underestimated for all variables, indicating that daily SLP cannot fully predict lower frequency variations. This is related to a shorter time persistence of the reconstructed fields, which leads to a lower variance when the fields are time filtered. Regarding the loss of variance on monthly scale, the variables form three groups with SLP, wind speed and precipitation showing an underestimation of not more than 30 %, humidity and cloud cover with 40% (January) to 60% (July) and T2M with only 10% (January) to 20% (July).

The relatively good performance of the T2M reconstruction based on the combination of monthly means reconstructed separated from daily anomalies (Sect. 2.2.3.4) indicates that a further improvement might be possible also for other variables if different scales are reconstructed separately. This seems to be important for, e.g. monthly mean humidity and cloudiness, where daily SLP is not very well suited to predict their variations on longer time-scales. The disadvantage of reconstructing low-frequency variations separately, e.g. using also different variables (proxies) for different scales as predictor (Moberg et al., 2005), is that many more analogs are needed than are usually present in a 50 yr period in contrast to daily analogs.

2.4.3.3 Bias

Regarding the deviation in mean (monthly sum for precipitation) of HiResAFF for the different variables (Sect. 2.3.3.3), an E–W dipole pattern can be seen in January for variables with a strong physical link to SLP - SLP, wind speeds and precipitation. This is also the case for cloudiness and temperature. In January, wind speed, precipitation and also cloudiness show negative bias for the western and central domain largely affected by the westerly flow while overestimation towards the E coincides with the transition to continental conditions.

The remarkable bias of both signs for SLP in July leads to a different latitudinal gradient in the pressure fields of the reconstruction compared to the model simulation. The reason for this large deviation in the reconstruction is unclear. Obviously, pressure fields in July are not adequately reconstructed according to the RCAO simulation driven by ERA40, although SLP is used as predictor. Together with the large gap regarding correlations in the surrogate approach compared to those of HiResAFF for July (Fig. 2-3), the hypothesis that discrepancies between observed SLP and simulated SLP seem to be model dependent, is further supported. A further investigation of this feature would however require an inter-model comparison which is beyond the scope of this paper.

In the case of winter T2M, a clear warm bias dominates over land whereas the Baltic Sea shows only a small bias compared to a cold bias over the North Sea and the NE-Atlantic. In summer, partly significant cold bias is reconstructed for continental regions in the SE but also N-Scandinavia. Humidity and cloud cover show spatially heterogeneous bias of both signs in July due to dominating small-to meso-scale processes. Precipitation shows mostly significantly too low precipitation amounts in July.

2.4.3.4 Reconstruction of frequency distributions and autocorrelation

From the results shown in Fig. 2-9, the ability of the AM to reconstruct realistic probability distribution of all variables is evident as a typical property of the AM method in general (Zorita and von Storch, 1999; Fernández and Sáenz, 2003).

In principle, the AM would also be able to reconstruct the observed probability distributions even if the predictor had no predictive skill at all, since the AM just re-orders the predictand data in time. Hence, the challenge for the analog-upscaling (or downscaling) is to achieve good temporal correlations between the reference and reconstructed variables and a realistic persistence in the reconstructed fields.

Owing to the memory/persistence in the climate system, a typical property of daily time series of atmospheric variables is their non-zero serial-correlation. While – dependent on the variable – consecutive days are not independent from each other, the AM used in the standard

approach (Sect. 2.2.3.2) does not take this persistence explicitly into account, since the analogs for two consecutive days are independently searched. Whether serial-correlation is still reconstructed by the AM fully depends on whether or to which extent the memory contained in the SLP predictor data is also related to the memory of the predictands like humidity or temperature, etc. As shown in Fig. 10a, realistic persistence is therefore only partly reconstructed by the AM for variables with a close link to daily SLP as predictor.

For T2M, the alternative approach of explicitly introducing persistence over four days (n-lag = 4, Sect. 2.2.3.3) when searching for analogs shows a clear improvement in the reconstructed autocorrelation. In this case, SLP is used to search for the best block of n-lag days, while the persistence of T2M stems from the memory contained in the block of consecutive days with length n. The disadvantage here is that the best analog for a given day is not necessarily contained in the best block of n days. As a consequence, the usage of n-lags >4 would lead to a decrease in the field correlation (Fig. 2-10b). Which approach is used depends in the end on the purpose of the study and the question of whether persistence of different variables is more important than to find the best analogs for single days.

2.4.4 Added-value vs. bias when using model fields as analogs

For the evaluation of the AM and the validation of HiResAFF, we chose the fields from the regional climate model RCAO (RCAX in T2M case) as reference. Using a leave-one-out approach for the reconstruction, skipping always the actual year from the analog pool, the fields are temporally independent but share the same physics/properties and model bias for the different variables. The principal added value of using fields from state-of-the-art RCMs as analogous fields relates to their physical consistency and the highly resolved regional to local features. Using the AM for upscaling, this study shows that a relatively sparse density of stations (proxies) can be used to predict corresponding atmospheric fields. The advantage of the AM compared to interpolation or regression techniques is that the fields themselves do not need to be reconstructed from the data – which would be impossible regarding physical consistency based on statistical methods.

However, as already mentioned, comparing the reconstruction with different observations, potential users of HiResAFF or similar reconstructions should be aware of the additional bias contained in the forcing fields, which stem from the used atmospheric fields of the ERA40 driven RCAO (Meier et al., 2011b) or RCA3 (Christensen et al., 2010) simulation. This model bias is principally independent from the bias caused by the AM shown in this study but will affect the reconstruction e.g. when being used as forcing data. As shown in Fig. 3 comparing the reconstruction in a surrogate climate (case Ref) with HiResAFF (case A), considerable deviations also in time are possible when linking observations to models driven by reanalysis data only at their boundaries and without spectral nudging (Sect. 2.4.1).

The chosen RCAO is a state-of-the-art RCM specially designed to interactively couple the air–sea–ice fluxes with the Baltic Sea ocean model. As shown by Meier et al. (2011b), the coupled ocean leads to a significant improvement of simulated winds over the Baltic Sea compared to an atmosphere-only version (RCA3). However, besides typical deviations in temperature and precipitation, etc., also the treatment of wind in different RCMs is important when using the reconstruction as forcing fields. As shown by Rockel and Woth (2007), RCMs tend to simulate generally too low wind speeds for higher percentiles when no gustiness correction is applied to the model output. As the used RCAO currently do not provide this correction, high wind speeds tend to be systematically underestimated already by the used fields, regardless of the AM's skill to reconstruct extreme wind speeds.

Based on the results shown in Fig. 2-3 and the discussion in Sect. 2.4.1, the analog-upscaling is always to some extent model dependent. In general, different models and settings from those used for HiResAFF can be used and the choice depends in the end on the users preferences. One aspect regarding the reconstruction of forcing fields, e.g. for ocean and ecosystem models, is related to the possibility of using the same RCM for the reconstructed fields and scenario runs for future climates (Meier et al., 2011a, 2012b, 2014). In this case, the atmospheric forcing remains consistent regarding the properties of the model (e.g. model bias etc.) throughout the whole time period. This might be an important advantage for detection and attribution studies related to ecosystem modelling.

2.5 Summary and conclusion

The AM used as nonlinear upscaling tool has been evaluated to reconstruct high-frequency variability of multivariate atmospheric fields on daily and monthly scale for a 50yr period. Based on up to 23 stations providing daily SLP as predictor, the AM is suitable to successfully reconstruct variables with a strong physical link to SLP, i.e. atmospheric fields of SLP and wind. For the wind reconstructions, the temporal correlations between HiResAFF and the reference simulation indicate a dependency on the intensity of the westerly flow. This means that the dominating large-scale circulation over the western domain yields higher reconstruction skills towards the NE-Atlantic and decreasing skill over the eastern and southern parts of the domain. The decrease in skill towards the east is most likely caused by the transition to more continental climate conditions with less influence of intense westerly winds and in contrast higher spatial variability. In order to successfully reconstruct atmospheric conditions off the coast and/or over complex topography, the AM needs more local predictors than for regions being better described by the large-scale circulation only.

This is also partly the case for precipitation. Reconstructed precipitation fields show a clear seasonal difference in correlations with very high skill during winter related to the dominating large-scale advective processes. The regionally lower skill during summertime may be attributed to local small scale convective processes which cannot or can only barely be captured by the large-scale SLP predictor field. Limitations within the RCAO simulation are a possible explanation for additional deviations due to not adequately resolved small-scale processes in the simulation, e.g. related to convection. The reader should be reminded here that Matulla et al. (2008) suggested different settings for the AM when reconstructing precipitation for downscaling. No such optimization is evaluated here to keep the different fields physically consistent.

For the reconstruction of cloudiness and relative humidity, daily SLP was also used as predictor. Due to the complex nature controlling the temporal and spatial variability of these two variables, only weak but still significant correlations between HiResAFF and the reference simulation are achieved over many regions. It should be noted that low reconstruction skills for these variables might also be caused by a different physical link in the model and in reality between SLP and these variables. The marked regional differences between land and ocean regarding correlation skills likely indicate that SLP is not simultaneously suitable to predict other variables for both surface types. The strong underestimation of variance in cloudiness and humidity on monthly scale indicates that daily SLP is not a suitable predictor in this case on longer time scales. The advantage of the AM is here restricted to the physical consistency of the fields, providing mostly satisfying correlations for both variables on monthly scale together with a realistic reproduction of probability distributions and their regional modifications represented in the regional climate simulation.

Due to the weak physical link between SLP and air temperature, monthly mean temperature fields were reconstructed using additionally 22 stations providing monthly mean temperatures as predictor. The idea of separating the reconstruction of different time scales using different predictors as in the case of T2M (Sect. 2.2.3.4) is similar to the approaches of Moberg et al. (2005) and Guiot et al. (2010) and might be used also for other variables or multi-proxies when applying the AM. In this case, however, two aspects need to be considered. First, a meaningful variable for the predictor is required e.g. to capture precipitation changes that are related to thermodynamic (in contrast to simply dynamic) processes (Matulla et al., 2008). Second, the strongly reduced number of available analogs should be kept in mind when searching for monthly or even seasonal patterns instead of daily analogs.

In the case of T2M reconstruction in this study, the size of the analog pool of monthly data is considerably reduced compared to the daily data. However, a first evaluation of the long-term trends and low-frequency variability shows a good agreement with long historical observations over the Baltic Sea region (Gustafsson et al., 2012) when searching for monthly analogs also in neighbouring months (M3 pool, Sect. 2.2.3.4). The high-frequency temperature anomalies reconstructed by daily SLP, which are added onto the time-interpolated monthly mean T2M, show seasonally different skill for correlation and variance. Introducing persistence over four days (n-lag = 4) in the analog search considerably improves the replication of serial correlation in daily temperatures, which is important for, e.g. the forcing of ecosystem (biochemical) models. Using daily near-surface temperature from model grid points as pseudo-predictors, the AM also yields very high reconstruction skills for near-surface temperature fields (not shown). Hence, digitized and homogenized daily historical near-surface temperature observations will be needed as predictor in subsequent studies to further improve the daily temperature reconstruction.

From the evaluation of the 50 yr presented in this study, it can be concluded that the reconstructed dataset of HiResAFF and the AM used as nonlinear upscaling tool is able to realistically replicate the high-frequency variability on daily and, with the exception of humidity and cloudiness, also on monthly scale. The frequency distributions and temporal correlations of multiple meteorological variables are well reconstructed. On daily scale, SLP and wind provide high confidence in a realistic reconstruction of extreme values with a high temporal and spatial co-variability consistent to the reference fields. This is important, for example, for ocean and ecosystem models and regions with complex topography like the Baltic Sea. The reconstructed fields of near-surface temperature, relative humidity, cloudiness and precipitation show realistic statistical properties and physical consistency on a daily scale with increasing confidence in the monthly to seasonal correlations compared to the reference fields. The monthly and seasonal resolution provides reasonably high quality when used as meteorological forcing fields.

Based on the successful validation of the analog-upscaling for the 50-yr period in this study, the evaluation of the reconstruction will be extended back to 1850 in a following study in order to estimate the AM's ability to also reconstruct low-frequency multi-decadal variations predicted by daily SLP and monthly air temperature. As the number of stations has been already limited in this study, similar reconstruction skills are expected at least back to 1870, with increasing uncertainties till 1850 due to the reduced availability of daily SLP. First results for the Baltic Sea (Gustafsson et al., 2012; Meier et al., 2012b, 2014) indicate that the AM also realistically reconstructs long-term changes when HiResAFF is used to drive ecosystem models for the Baltic Sea extending back to 1850.

3 Long-term storm variations since 1850

Numerous studies based on historical documentary reports, long-term surge and high-tide data or pressure-based wind indices mostly do not suggest robust long-term changes in the storm climate over the Euro-Atlantic region since the late 19th century or longer. This is in sharp contrast to the novel 20th century reanalysis (20CR) showing strong significantly positive storm trends in this region since 1871. As such a trend would have high impacts for instance on coastal communities, near- and offshore infrastructures, forestry or ecosystems etc., the extended reconstruction of HiResAFF (chapter 2) might serve as an important alternative to estimate homogeneous and physical consistent long-term information on pressure and wind fields. In contrast to 20CR, the reconstruction makes use of only a limited but instead rather constant number of pressure observations. A comparison with other historical storm information is hence needed to test whether the chosen settings and predictors of HiResAFF adequately capture low-frequency variations over different regions.

In chapter 2, the reconstruction of HiResAFF through analog-upscaling has been shown to provide realistic climatological properties with physically consistent variable- and region-specific frequency distributions and realistic temporal correlation and variance on daily to monthly scale. This chapter provides an analysis of reconstructed low-frequency variations and potential trends in wind and pressure fields since 1850 in comparison to existing long-term wind and pressure information. As the novel 20CR aims to provide comparable atmospheric fields since 1871, the results regarding low-frequency variations and trends are compared with those of HiResAFF. In addition, other pressure-based wind indices over Northern Europe are analysed regarding their annual and decadal co-variability. To gain a conclusive view on potential long-term trends in storminess, results are discussed in the context of a thorough literature review of available historical studies on long-term trends.

3.1 Introduction

A strong and systematic interest into long-term variations or trends in the wind climate over the Euro-Atlantic region started only about 20 years ago. A clear change towards a roughened wind and wave climate over the NE-Atlantic and North Sea from the 1970s to 1990s (c.f. Carter and Draper, 1988; Hogben, 1994) raised public concerns about a potential impact of global warming on the storm climate (Schmidt and von Storch, 1993). Little was, however, known about long-term storm variations and possible links to internal or external drivers.

Variations in European storminess are closely linked to the baroclinic activity over the North Atlantic and the strength of the zonal pressure gradient between Iceland and the Azores. Variations of this gradient represent the dominant mode of lower and mid-tropospheric pressure variability over the North Atlantic. These variations are described by the North Atlantic Oscillation (NAO) which explains about one third of winter storm variations over Northern Europe (Hurrel, 1995). Also long-term variations of annual 99th percentile geostrophic wind speeds over Western Europe (Wang et al., 2011) or storm indices over the Baltic Sea region (Bärring and Fortuniak, 2009) agree with variations of the NAO.

The link to the NAO is however not stationary over time (Matulla et al., 2007) with a higher correlation to the NAO in recent decades (Alexander et al., 2005) and a rather weak link previously (Alexandersson et al., 1998). A strong pressure gradient (NAO+) is usually connected with high baroclinic activity, a strong westerly flow and increased storminess over Northern Europe while weak gradients (NAO-) lead to a displacement of storm tracks further

south with calm or even easterly wind conditions over Northern Europe (Hurrel et al., 2003). Nevertheless, severe cyclones can also develop during negative phases of the NAO (Pinto et al., 2009). Also an increased frequency of atmospheric blocking through quasi-stationary high pressure systems over the Euro-Atlantic region can interfere with the zonal flow leading to strong seasonal anomalies on regional scale (Rimbu and Lohmann, 2011). Including also variations in the strength and westward extension of the thermally driven Siberian High in the winter half-year, different studies on long-term variations in atmospheric circulation and weather types indicate an increased frequency of anticyclonic and westerly types over the last 200 (Omstedt et al., 2004) to 300 years (Jacobeit et al., 2003, Hurrell and Folland, 2002).

To estimate whether the strong increase at the end of the 20th century falls within the range of large decadal natural variations, longer timescales need to be considered. It became clear that the shortness and inhomogeneity of most wind records even in recent decades (cf. Lindenberg et al., 2012) compromise studies on long-term trends in the wind climate. The project *Waves and Storms in the North Atlantic* (WASA, 1998) therefore suggested the use of pressure-based storm indices owing to the pressure's lower sensitivity to non-climatic influences and measurement errors. Recent studies confirm a reasonable to high informational value of pressure-based single station proxies (Krueger and von Storch, 2012) or high annual percentiles of geostrophic wind speeds calculated from triplets of station pressure (Krueger and von Storch, 2011).

Different pressure-based proxies have been used since then indicating no robust long-term trends over the Euro-Atlantic region since the 1880s or earlier (cf. Schmidt and von Storch, 1993; Alexandersson et al., 2000; Matulla et al., 2007; Hanna et al., 2008; Bärring and Fortuniak, 2009; Wang et al. 2009a; 2011). Large decadal variations rather than robust trends are also confirmed by multiple wind and storm related historical records reaching back to the 15th century for the southern North Sea (de Kraker, 1999). Also the so far longest continuous wind observations do not indicate any systematic changes over centuries in the Dublin region since 1715 (Sweeney, 2000). These results seem to support studies of externally forced coupled atmosphere-ocean general circulation model (AOGCM) simulations (e.g. ECHO-G) for the last millennium indicating quasi-stationary wind and cyclone statistics over the northern hemisphere (Fischer-Bruns et al., 2005; Xia et al., 2012).

Nevertheless, studies on changes of storm tracks over the North Atlantic suggest a clear northeast shift since around the 1970s (Sickmöller et al., 1997; Chang and Fu, 2002; Wang et al., 2006) together with a strong increase in the number of deep lows in winter and spring (Lehmann et al., 2011). Based on gridded reanalysis data for the last 40-60 years, studies indicate positive trends for the wind climate and strong cyclone activity north of around 55°-60°N over the North Atlantic while regions further south exhibit negative trends (Gulev et al., 2001; McCabe et al., 2001; Trigo et al., 2006; Raible et al., 2008). The spatial change in storm tracks over the North Atlantic appears to be consistent with scenario simulations until 2100 under increased greenhouse gas concentrations (cf. Ulbrich et al., 2009). A first study for the wind and wave climate over the North Atlantic suggests that combined external and natural forcing can explain changes in the period 1950-2000 while external forcing had less likely an influence on the first 50 years of the 20th century (Wang et al., 2009b).

While AOGCM simulations for the last millennium, historical observations since the 15th century and pressure-based storm indices since around 1800 suggest no long-term trends in the wind climate over the North Atlantic, the novel 20th Century Reanalysis data (20CR, Compo et al., 2011) points towards significant positive trends in annual storminess over the NE-Atlantic and North Sea since 1871 (Donat et al., 2011). A first study shows that the

inconsistency of 20CR trends with pressure-based storm indices over the NE-Atlantic is very likely caused by clear changes in the amount of assimilated data into 20CR over time (Krueger et al., 2013). A relative good agreement of 20CR with storm records in Zürich since 1871 (Brönnimann et al., 2012) might indirectly support this hypothesis as the availability of data is especially very low over sea areas back in time and less so over central Europe.

The introduction of spurious trends into reanalysis models through changes of the number, type or quality of assimilated observations is not new (cf. Kistler et al., 2001; Bengtsson et al., 2004) leading to well known artefacts/errors over data sparse regions like the southern hemisphere (cf. Hines et al., 2000; Bromwitch et al., 2007). A reduced reliability of 20CR is also stressed by Compo et al. (2011) for the last 50 years over the southern hemisphere and by Fergusson and Villarini (2012) over the US. Suggestions to produce reanalysis products from a reduced but constant amount of assimilated data (Thorne and Vose, 2010) has not been taken up so far.

The datasets of both, 20CR and HiResAFF, aim to provide physically consistent atmospheric fields on longer timescales making use of mainly daily historical pressure data. While 20CR assimilates all available records over time, the approach chosen to reconstruct HiResAFF in chapter 2 aims to minimize potential (so far unknown) effects possibly introduced by the considerably changing number of pressure observations over time. As HiResAFF is a new reconstruction and 20CR appears to be inconsistent with observations according to a first long-term evaluation (Krueger et al., 2013), some well established pressure-based wind indices over the North Sea, NE-Atlantic and Baltic Sea are used as robust and independent measures for variations of annual wind statistics since the 1880s or earlier.

In order to analyse the long-term wind climate and evaluate the consistency of HiResAFF with available long-term datasets, the following chapter focuses on the co-variability of annual wind and storm statistics between HiResAFF, 20CR and different wind indices over Northern Europe. Centennial trends of annual median and 95th percentile wind speeds of HiResAFF are compared with those derived from 20CR since 1871. An alternative test reconstruction of HiResAFF with a very small number of predictors is also considered to estimate the impact of increased uncertainty on low-frequency variations and long-term trends in remote regions over the NE-Atlantic. Thereby it is tested whether long-term changes in storminess exist in different datasets and whether they are consistent with other studies. To give a profound and more definite answer to the question about the existence of long-term changes in the wind climate, results of the long-term trend analysis are discussed in the context of a synthesized literature review of available mostly regional studies on the historical wind climate.

3.2 Data and Methods

3.2.1 Extended reconstruction of HiResAFF 1850-2009

The analog-reconstruction of HiResAFF is extended based on the same method and predictors as published in Schenk and Zorita (2012) in chapter 2 to cover the period 1850-2009. Details about the used daily pressure observations are listed in Tab. 3-1. Figure 3-1 illustrates the temporal data coverage for the different stations. Details about the historical data and applied homogenizations can be found in Ansell et al. (2006) and Klein Tank et al. (2002). Updates from different institutions listed in Tab. 3-1 have been compared for overlapping periods mostly excluding mountain stations due to large errors relative to neighbouring stations.

The availability of station data decreased slightly with the breakdown of former Soviet countries after 1990. A stronger decline in the station numbers has to be considered before around 1874 with only six to nine stations during the first year 1850 (Fig. 3-1). A first estimate of uncertainties introduced through using different numbers and locations of predictors has been applied in chapter 2-3.2.3 in terms of correlation between reconstructed and simulated wind speeds on the daily and monthly scale (Fig. 2-5). The test case with only six stations as in Fig. 2-5 (c3) is now reconstructed with real pressure observations for the period 1850-2009, representing the worst case situation during parts of 1850 as lowest confidence estimate. Due to the fact that these stations are located in the central domain, a comparison of this test case (Test_N1850) with HiResAFF is applied regarding co-variability and long-term trends over remote regions relative to the predictors used around the Baltic Sea.

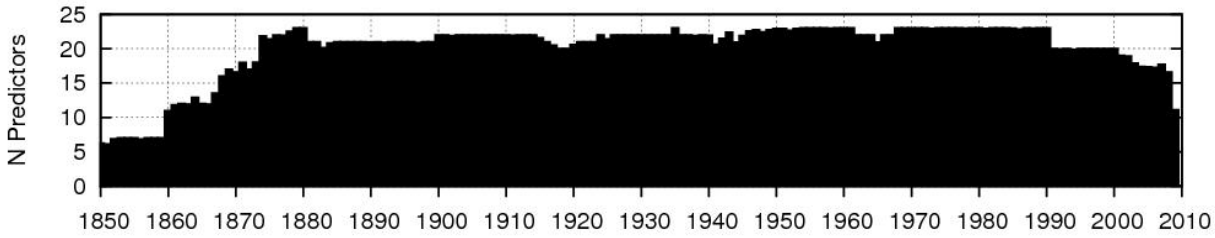

Fig. 3-1: Temporal changes in the number of available stations (total: N=23) for the reconstruction of HiResAFF 1850-2009. Parts of 1850 consist of only N=6 (case c3 in Fig. 2-5) used as Test_N1850 in the following.

Table 3-1: Station of daily sea-level pressure (SLP) used for the reconstruction of HiResAFF 1850-2009. The second source refers to the update. Source [1] = EMULATE (Ansell et al., 2006), [2] = ECA&D (Klein Tank et al., 2002), SMHI = Swedish Meteorological Hydrological Institute, FMI = Finish Meteorological Institute, Jena = University of Jena, ZAMG = Zentralanstalt für Meteorologie und Geodynamik, Vienna/Austria. Missing values for each station in % for the whole period 1850-01-01 until 2009-11-30. See Fig. 2-2 for a geographic map of station locations.

Location (name)	WMO Identifier	Lat. [° N]	Lon. [° E]	Height [m]	Start [year]	End [year]	Added [year]	Misval. [%]	Sources
Bergen (Flesland)	1317	60.38	5.33	36	1868	2002	2009	11.27	1, 2
Bodø	1152	67.26	14.43	13	1868	1994	2009	23.17	1, 2
deBilt	6260	52.10	5.18	2	1850	2009		0.02	2
Gothenburg	2526	55.70	11.98	155	1860	2002	2009	6.28	1, 2
Hammerodde	6193	55.30	14.78	11	1874	2009		16.75	2
Haparanda	2196	65.82	24.13	6	1860	2002	2008	6.78	1, SMHI
Harnosand	2365	62.61	17.93	6	1860	1995	2008	6.77	1, SMHI
Helsinki	2974	60.17	24.95	56	1850	2001	2009	4.16	1, FMI
Hohenpeissenberg	10962	47.80	11.02	977	1850	2002	2009	4.28	1, 2
Jena	10962	50.93	11.58	155	1850	2000	2009	3.65	1, Jena
Kiev	33345	50.40	30.45	179	1850	1990	n.a.	25.53	1
Kremsmünster	11012	48.05	14.13	383	1876	2009		18.39	ZAMG
Lund	2627	55.70	13.20	73	1864	2001	2008	9.38	1, SMHI
Nordby	6080	55.43	8.40	29	1874	2008		15.62	2
Oksoyfyr	1448	58.07	8.05	8	1870	2002	2009	12.62	1, 2
Riga	26422	56.81	23.89	13	1850	1990	n.a.	21.77	1
Stockholm	2485	59.33	18.05	44	1850	1998	2008	0.47	1, SMHI
Saint Petersburg	26063	59.93	27.96	6	1850	2000	n.a.	5.48	1
Torshavn	6011	62.02	-6.77	55	1874	2009		15.53	2
Vardø	1098	70.36	31.10	15	1861	2002	2009	43.34	1, 2
Vestervig	ECA107	56.77	8.32	19	1874	2009		15.11	2
Visby	2591	57.63	18.28	47	1860	2002	2009	6.37	1, 2
Vilnius	26730	54.68	25.30	156	1850	1990	n.a.	27.45	1

3.2.2 Analysis of long-term variations and trends in storminess

In a first step, the large-scale co-variability of storminess of HiResAFF and 20CR is analysed based on the leading modes of annual wind and storm variability over Northern Europe by means of Empirical Orthogonal Function (EOF) analysis. The EOFs are calculated from spatially weighted anomaly fields of annual 50^{th} and 95^{th} percentile wind speeds over the period 1871-2008. Corresponding time series of the Principal Components (PC) are compared between both datasets. Their physical meaning is evaluated by projecting the PC time series onto anomaly fields of annual 5^{th} percentile (low pressure) sea-level-pressure fields. Changes in the annual number of days with a field minimum sea-level-pressure falling below 980 hPa (N<980hPa) are counted. Because observations since the 1970s indicate also a NE shift of deep lows towards Northern Europe (e.g. Lehmann et al., 2011), decadal variations of the annual area affected by deep lows [km² per year] are additionally analysed. Although correlations per se do not necessarily reflect physical processes, meteorological considerations together with a high statistical correlation provide a sound basis to make inferences about the link of deep lows and storminess. Spearman rank correlations are used here to quantify the link of storm variations with variations of N980 and their spatial coverage.

In a second step, annual and decadal storm time series of HiResAFF and 20CR are analysed for different regions. The 20CR dataset provides an ensemble mean together with 56 ensemble members that represent plausible reconstructions compatible with the assimilated data given the present data and model uncertainties. The ensemble mean is presented together with the 90% model spread derived from these 56 ensemble time series. For HiResAFF, Test_N1850 serves as low confidence version of HiResAFF to evaluate potential effects of a reduced number of predictor stations on remote regions. The degree of co-variability on the regional scale is estimated through correlation maps of annual and decadal storm variations. Statistical significance ($p<0.05$) for annual correlations is inferred through local 2-sided t-tests. In case of running decadal mean storm variations, significance levels are estimated through block bootstrapping to account for the high serial autocorrelation of the low-pass filtered time series. A block length of nine is used here (n-lag for $r< 0.2$). Field significance ($p<0.05$) is estimated as in chapter 2.2.5 through block bootstrapping with fixed time patterns.

To quantify potential long-term trends of annual 50^{th}, 95^{th} and 99^{th} percentile wind speeds for different regions, slope parameters are estimated through linear regression. A significant deviation of local slope parameters from 0 is estimated by a parametric 2-sided t-test ($p<0.05$). The results for HiResAFF are compared with those of 20CR. For the latter, the ensemble mean is chosen. To estimate the uncertainty of trends from the 56 ensemble members, the absolute spread between the ensemble members with the smallest (member 6) and largest (member 17) centennial trends are presented. For HiResAFF, the differences of centennial trends with Test_N1850 are used to estimate potential regional deviations in remote regions. For a better comparability, all trends are given as relative change in per cent per century with respect to the long-term mean.

In a next step, pressure-based storm indices are used as an additional independent source to analyse long-term storm statics over different regions. First, wind indices of annual 95^{th} percentile geostrophic wind speeds from different triangles over the NE-Atlantic and North Sea are used. The time series by Alexandersson et al. (2000) have been updated to 1881-2004 and were recently also used by Krueger et al. (2013). One-point-correlation maps are calculated to estimate the spatial extent of co-variability between wind indices and annual 95^{th} percentile wind fields of HiResAFF. The same analysis is applied for similar geostrophic wind indices over the German Bight (Schmidt and von Storch, 1993) updated to 1879-2012 and Finland (Suvilampi, 2009) 1884–2007. Also HiResAFF's co-variability with different

annual single-station pressure indices of Stockholm is briefly assessed for the period 1850-2005 (Bärring and Fortuniak, 2009). For geostrophic wind indices, a parametric 2-sided t-test is used to test whether correlations are significantly different from 0 with $p<0.05$. For the single-station indices of Stockholm, Spearman rank correlations are used to allow for a better comparison of the co-variability with different proxies sharing different frequency distributions. For correlations of decadal running mean storminess, significance levels are estimated by block-bootstrapping as before to account for the high serial (auto)correlation of time series.

In addition to the previous analysis, the inconsistencies of 20CR's low-frequency variations and trends in storminess are briefly assessed for the more regional scale over Northern Europe. First, correlation maps between 20CR and the averaged time series of the triangles over the NE-Atlantic by Krueger et al. (2013) are calculated for the period before and after 1940 and 1881–2004, respectively. To separate the annual co-variability from the long-term trends, the correlation maps are additionally calculated after subtracting the trends. For HiResAFF, the correlation maps of decadal running mean storminess with 20CR are also compared with the detrended results. Together with the regional analysis of time series before, the regions of main inconsistencies regarding long-term trends in 20CR are discussed.

The results regarding low-frequency variations and trends of Northern European storminess are discussed in the context of other historical studies. For this purpose, regional storminess trends of available observational and proxy studies are synthesized into one figure, highlighting the (non-)existence of trends dependent on the time period and region. All studies containing keywords like *wind, storm or storminess* in the online database of *Science Direct* (www.sciencedirect.com) since 1990 are considered (as of November 2012).

3.3 Results

3.3.1 Comparison of large-scale storm variations with 20th Century Reanalysis

As a first comparison between HiResAFF and 20CR, the overall agreement of large-scale wind variations over Northern Europe is analysed for the period 1871–2008. An EOF analysis of annual mean and annual 95th percentile wind speeds between HiResAFF and 20CR (m06, m08, m17) yields very similar patterns for the first two leading modes and thus only those of HiResAFF are shown in Fig. 3-2 for storminess. The major differences in the EOFs relate to stronger loadings in case of 20CR owing to slightly higher amplitudes in the time series of Principal Components (PC) of the 1st and 2nd EOF compared to HiResAFF. Higher EOFs are not considered here, because very similar eigenvalues point towards not well separated (degenerated) modes (von Storch and Zwiers, 1999).

The first leading pattern explains about 34.5% of the storm variability (20CR ~36%). The centre of variability is located along the Norwegian coast. Projecting the time series of PC1 in Fig. 3-2 onto anomalies of annual 5th percentile SLP fields yields a clear zonal dipole with positive correlations in the south and negative correlations in the north (contour lines in Fig. 3-2). This clearly demonstrates that the dominant mode of storm variability relates to strong/weak zonal gradients in the pressure anomaly field. As a result, the PC1 time series in Fig. 3-2 clearly reflects the main variations of storminess over northern Europe.

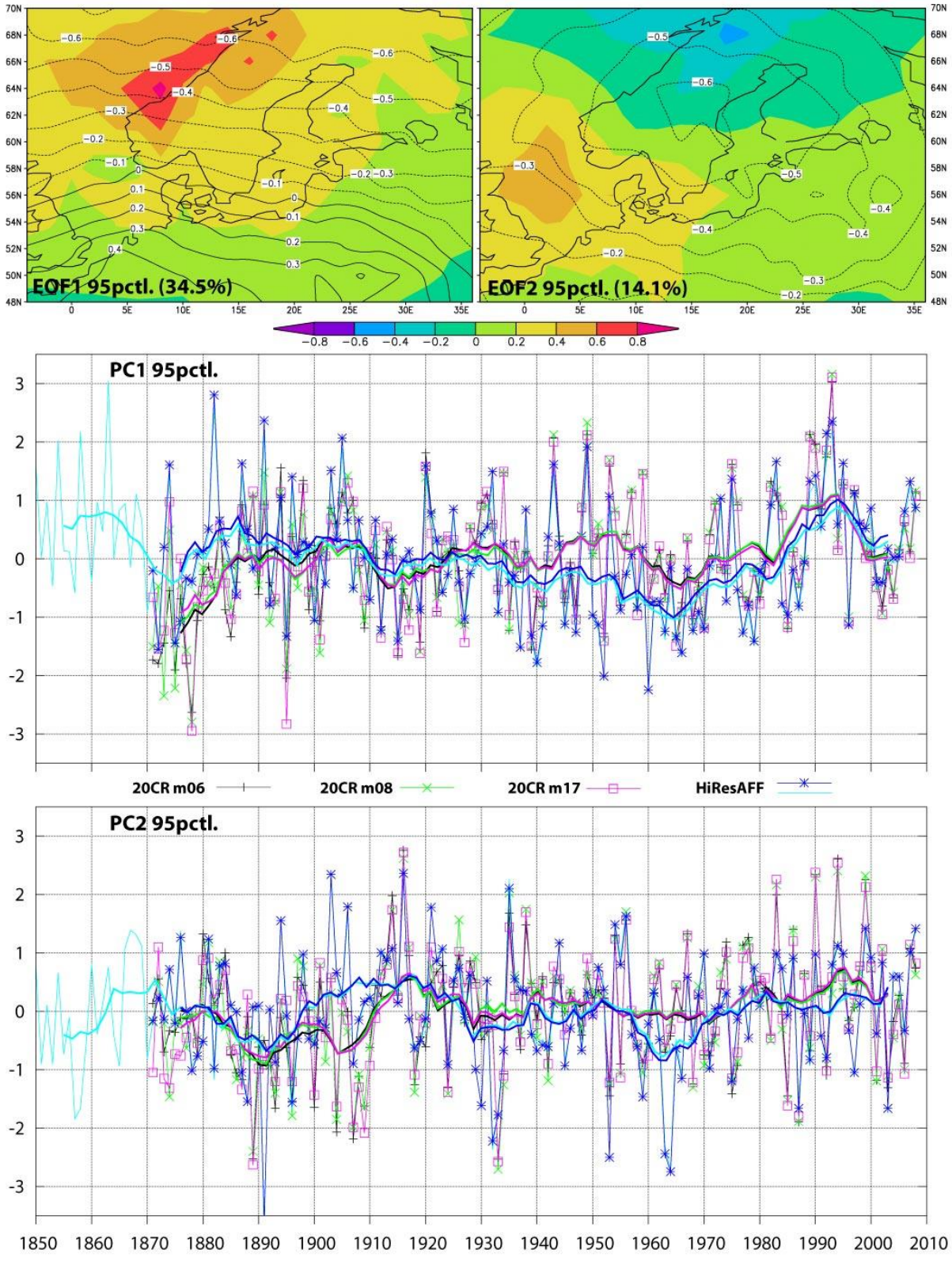

Fig. 3-2: 1st and 2nd EOF of the 95th annual percentile wind speed variability over Northern Europe. A) Loadings of the 1st and 2nd EOF pattern (shaded colours) and correlations (contours) of corresponding PC1 and PC2 with anomalies of the annual 5th percentile SLP fields of HiResAFF, B) time series of the 1st PC and C) 2nd PC of 95th percentile wind speeds of HiResAFF compared to those derived from three ensemble members of 20CR. 11year running means are used to highlight low-frequency variations. With exception of the extended HiResAFF 1850-2008 (light blue), the comparison refers to the period 1871-2008.

The annual storm variations of HiResAFF are on average in good agreement with PC1 derived from 20CR regarding magnitudes and phase ($r \sim 0.7$, $p < 0.05$). Differences highlighted by running decadal means exist in the 1950s to 1960s and before around 1890. The extended PC1 of HiResAFF since 1850 (light blue) shows a similar stormy period in the 1870s as in the 1990s. While PC1 of HiResAFF shows no trend since 1850 or 1871, 20CR indicates a significant increase of 0.6 to 0.7 standard deviations per century since 1871.

The second mode of storm variability (14.1%) (20CR ~11%) shows a dipole with high loadings over the North Sea and northern Scandinavia. The regression of the PC2 time series with the annual 5th percentile SLP fields displays a monopole structure (contour lines) with high negative correlations over central Scandinavia. The 2nd mode of storm variability is hence associated with variations of low pressure centres over Scandinavia. The negative correlation of PC2 with the SLP field shows that very deep lows (very small SLP values) are related to very high wind speeds (positive anomalies in PC2) and vice versa.

The co-variability between the PC2 time series on decadal time scales is in general high except for two outstanding periods. While HiResAFF indicates a very small impact of deep lows in PC2 during the very calm storm period of the 1970s (PC1), 20CR remains on average conditions. A second period with clear differences exists in the 1910s. On the annual scale, PC2 of HiResAFF shares a correlation of ~0.55 ($p < 0.05$) with 20CR. While significant positive trends in 20CR (~0.5 to ~0.6 stdev/100 years) indicate an increase of storm variability through deep lows, HiResAFF shows only a weak and non-significant increase since 1871 (+0.08) or 1850 (+0.11).

3.3.2 Variations in the number of deep lows (N<980 hPa) since 1850

Investigations based on reanalysis products related to NCEP show a pronounced increase in the number of deep cyclones over the North Atlantic together with a NE-ward shift of storm tracks since the 1970s (e.g. Lehmann et al., 2011). The time evolution of PC2 of HiResAFF in Fig. 3-2 reflects an increasing effect of deep cyclones onto storm variations since the 1970s albeit no robust trend exists on longer timescales. Within this context, the annual number of days when the minimum of the pressure field falls below 980 hPa (N<980 hPa) in the period 1850-2008 is counted. The resulting time series of the annual number of deep lows (Fig. 3-3) clearly reflects the storminess time series of PC1 with very high cyclone numbers in the 1870s, followed by a rather monotonic decrease to very low numbers in the 1970s and a strong increase towards the end of the 20th century. While the annual mean N<980 hPa is around 30.2 (median=30.7; sigma=9.77), extraordinary high numbers fall within the most prominent storm decades with n=60 in 1990 followed by n=55 in 1863 with a minimum of only n=8 in 1963.

The rank correlation of the N<980 time series with PC1 amounts to r=0.64 over the period 1850-2008. The statistical relationship between N<980 and PC2 is also reflected by a rank correlation of r=0.43. The annual N<980 does not exhibit any significant trend at the 5% level with an increase of 0.96 deep lows per century since 1871 and a decrease of -2.01 lows since 1850. To get spatially more robust information about the impact of deep lows on Northern Europe, the field mean of the area-weighted annual extension of N<980 [10^6 km²/year] is depicted in Fig. 3-3 as running decadal mean area. Here, the spatial extension of the area affected by deep lows (mean=11.3 Mio. km²/a) is now much higher at the end of the 20th century compared to the 1870s, although also no long-term trend exists mainly due to very low values around 1940-80.

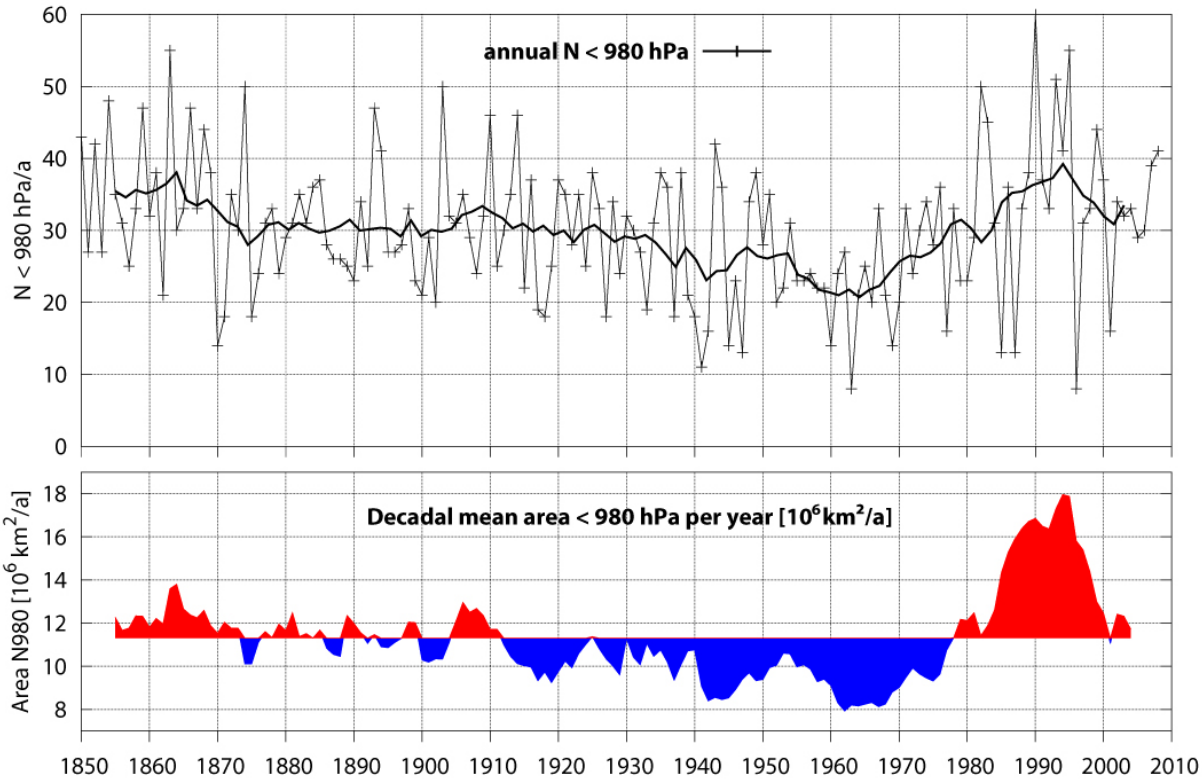

Fig. 3-3: Annual number of deep lows (N<980 hPa) over Northern Europe and 11yr running mean of the annual area affected by deep lows over the period 1850-2008.

3.3.3 Regional scale variations in storminess compared to 20CR

Figure 3-4 compares running decadal mean time series as anomalies relative to the mean 1960–1991 of annual 95th percentile wind speeds between HiResAFF (blue) and 20CR. In case of 20CR, 56 time series are calculated and the ensemble mean (green) together with the 90% ensemble spread (grey) are displayed. For HiResAFF, the uncertainty cannot be realistically estimated over the whole time series due to changes in the number and location of stations, mostly pronounced before 1874 (Tab. 3-1). Instead, a test case representing a reconstruction using only six stations as worst situation during parts of 1850 (Test_N1850) is depicted (light blue) as an alternative for uncertainty estimation.

With exception of the central to Eastern Europe, all regional time series of HiResAFF and 20CR start to deviate prior to the 1970s with at least some decadal co-variability over the North Sea and Central Scandinavia until around the 1930s. Due to the fact that the data are not normalized, 20CR shows generally stronger decadal variations. A strong downward trend in 20CR before 1900 in all regions is inconsistent with HiResAFF. Time series of Test_N1850 show higher decadal variations with mostly reasonable co-variability with HiResAFF over the central and eastern domain. Larger deviations exist most notably over the western domain where no predictors have been used in the test reconstruction.

To quantify the degree of co-variability of annual and decadal scale storminess of HiResAFF and 20CR for different regions, correlation maps are presented in Fig. 3-5. On inter-annual timescales, 80% of local grid points show statistical significant positive correlations at the 5% level with a field median correlation of 0.40 (m08). Local correlations are lower over large sea-areas of the Baltic Sea and North Sea with non-significant correlations over parts of the

NE-Atlantic and the SE. In contrast to inter-annual variations, correlations for decadal running mean storminess show a spatially more heterogeneous picture, reflected also by the regional time series in Fig. 3-4. On decadal time scales, a good agreement exists partly over land areas. No significant correlation exists over the North Sea and Norwegian Sea and the eastern Baltic Sea. The field median correlation amounts to only 0.32 with 60% local grid points showing statistically significant values (compared to 80% using 2sided t-test). The amount of locally significant grid points is globally significant ($p<0.05$) according to block bootstrapping. This result is however close to the threshold value of the 95% confidence level being 56.7%.

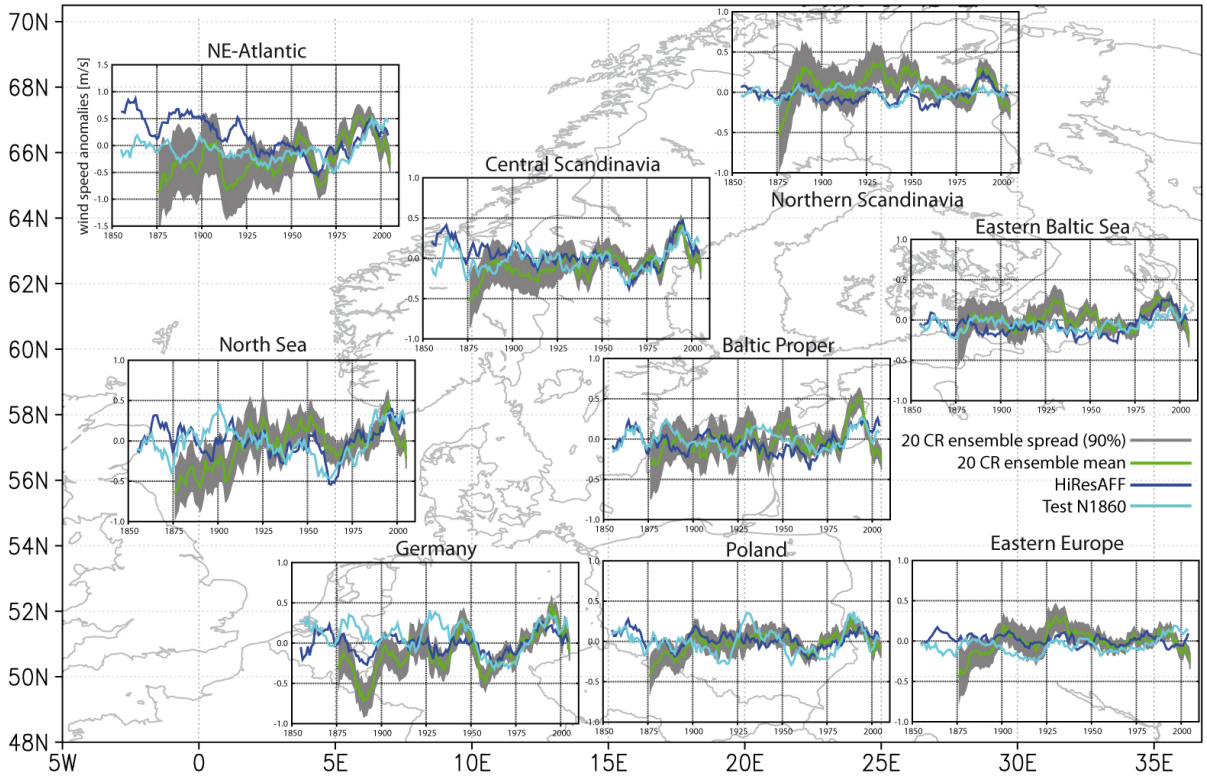

Fig. 3-4: Running decadal variations of annual 95th percentile wind speeds of HiResAFF (blue) 1850-2008 and 20CR 1871-2010 as anomalies relative to the mean of 1960-91. For 20CR, the ensemble mean (green) is presented together with the 90% model spread from the 56 ensemble members (gray). Test_N1850 (light blue) represents the reconstruction with only six predictors.

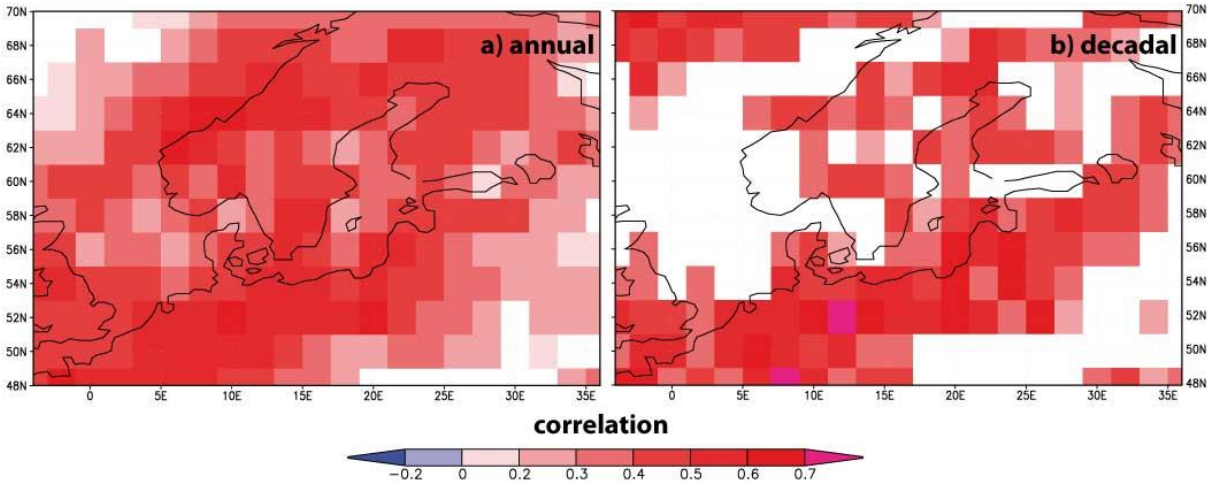

Fig. 3-5: Correlation of annual 95th percentile wind speeds between HiResAFF and 20CR (m08) on a) annual and b) decadal scale for the period 1871-2008 (non-significant ($p<0.05$) in white).

3.3.4 Long-term wind and storminess trends compared to 20CR

Based on 50th, 95th and 99th annual percentile wind speeds, linear trends are estimated for HiResAFF and 20CR over the period 1871–2008. For a better comparison, trends in Fig. 3-6 are given as change per century in percent relative to the long-term means. Grid points where no significant trends ($p<0.05$) exist are masked in white. To estimate the range of centennial trends of the 56 ensemble members, slope parameters are estimated first from all members before the ensemble mean trend in Fig. 3-6 is estimated.

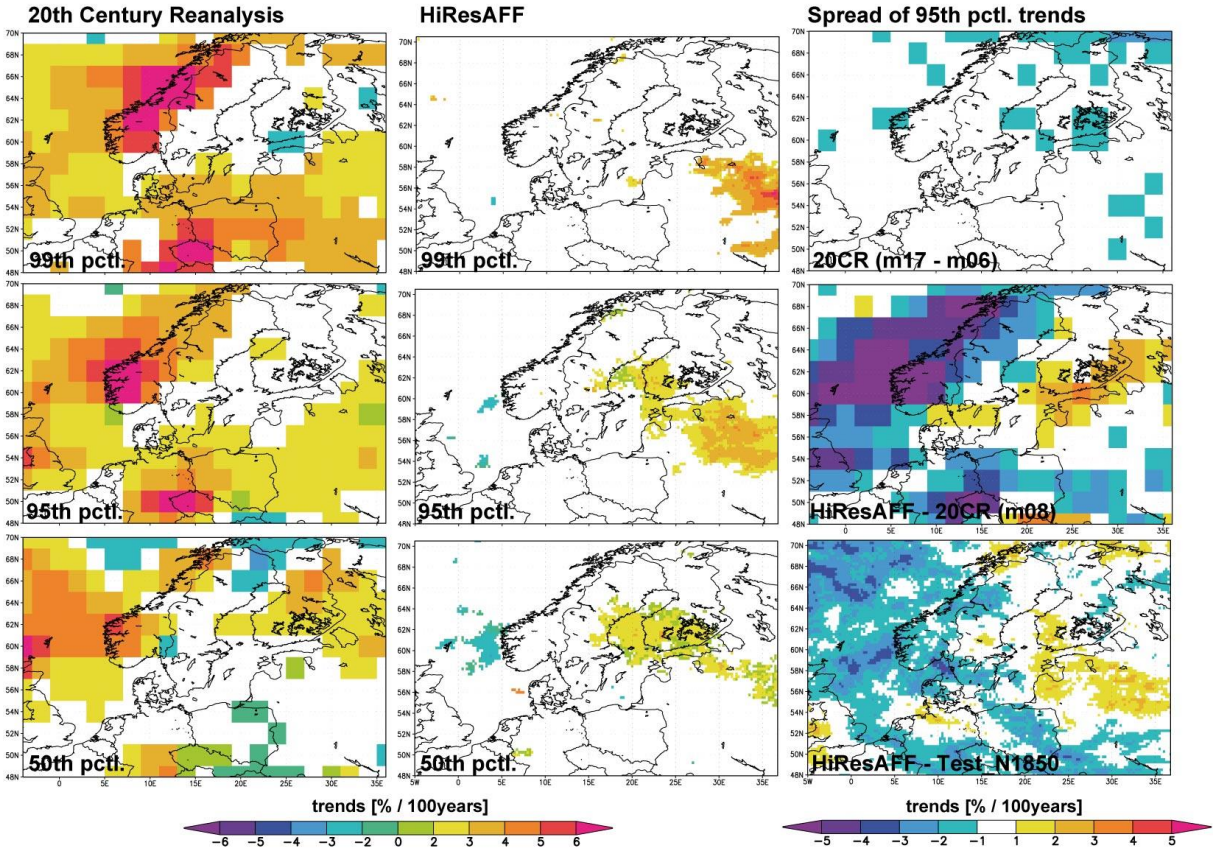

Fig. 3-6: Comparison of long-term trends [%/100 years, white: non-significant trends ($p < 0.05$)] of annual 50th, 95th and 99th percentile wind speeds for 20CR ensemble mean trends (column 1) and HiResAFF (column 2) over the period 1871-2008. Column3: Internal spread of 20CR (top) for trends of m17-m06 and difference of HiResAFF-Test_N1850 (bottom) with trend differences between HiResAFF-20CR (m08) (middle).

20CR shows strong positive trends (>7%/century) for high annual percentiles along the Norwegian coast with clear positive trends over the NE-Atlantic, North Sea and central Eastern Europe. For the annual median wind speeds, less pronounced but significant positive trends exist mainly north of 60°N over the NE-Atlantic and parts of NE-Europe. HiResAFF shows mostly opposed but insignificant negative trends over the North Sea and NE-Atlantic with weak significantly positive trends over isolated areas in the east. Test_N1850 yields weak positive trends over remote regions over the NE-Atlantic (not shown). The opposing trends are reflected by the trend spread between Test_N1850 and HiResAFF with a difference of up to 3%/century in remote regions (Fig. 3-6, right, bottom). The internal model spread of 20CR is rather small. The field median of 95th percentile trends of 20CR varies between +0.15 m/s (m17) and +0.20 m/s (m6) with an ensemble mean trend of +0.18 m/s (m8) per century. Subtracting the trends of m17 from m6 yields a maximum spread of only 3% varying mostly within ±1% (Fig. 3-6, right top).

As indicated by the 90% model spread in Fig. 3-4, the large differences result mainly from very low storminess in the early period which is most pronounced over the NE-Atlantic and not reflected in HiResAFF. As a result, subtracting the 95th percentile storm trends of 20CR (m8) from HiResAFF yields very large differences mainly over the NE-Atlantic with up to 8% along the Norwegian coast (Fig. 3-6, right middle). The non-significant correlations over sea-areas in Fig. 3-5 and the large discrepancies regarding long-term trends in the same region point towards clear inconsistencies, especially over open sea (see chapter 3.3.7).

3.3.5 Long-term storm variations compared to regional storm indices

North Sea and Norwegian Sea since 1881

First, one-point-correlations maps are calculated between time series annual mean and 95th annual percentile geostrophic wind speeds derived from different station triplets over the Norwegian and North Sea (updated after Alexandersson et al., 2000) and annual mean and 95th annual percentile wind speeds from HiResAFF. The level of agreement and spatial extension differs considerably for different triangles with annual means showing generally a higher agreement extended to a wider area around the respective triangle. The very good agreement with mean and extreme annual wind statistics over the British Isles (T6) and the Norwegian Sea (T3) confirm a realistic reconstruction of HiResAFF in this region where little predictor stations have been available. T5 shows a poor agreement for annual storminess within the triangle over the North Sea but a better agreement in surrounding areas while T10 shows a low agreement for storminess. The results are very similar when detrended data is used in case of HiResAFF. In contrast, a clear improvement is achieved after removing trends in case of 20CR (see chapter 3.3.7). The geostrophic indices and HiResAFF show both a slight negative tendency with in contrast significant positive trends in 20CR (Donat et al., 2011, Krueger et al., 2013).

German Bight since 1879

A comparison of annual 95th percentile geostrophic wind speeds derived from stations around the German Bight (updated after Schmidt and von Storch, 1993) with reconstructed percentiles of HiResAFF and 20CR is shown in Fig. 3-8 for the period since 1879. Time series of HiResAFF represent the spatial average of the 2°x2° area of the closest grid point of 20CR (54°N; 8°E) for the centre of the triangle. The time series are normalized with respect to the mean divided by the standard deviation of 1958–2005. A one-point-correlation map on the annual scale with HiResAFF is additionally shown in Fig. 3-9.

With the exception of the missing calm period of the 1970s in 20CR, HiResAFF and 20CR show a very good agreement until the 1930s. The geostrophic wind index shows a strong up-tick during the period 1940–50 that is not indicated by 20CR and HiResAFF. The increased storminess around 1920 is captured by the index and 20CR but not HiResAFF. While the index remains on high levels prior to 1930, HiResAFF indicates average conditions and 20CR shows below average values. As a result, the index exhibits a significant negative trend (-0.72 m/s or -3% per 100 years) with insignificant positive rates for HiResAFF and 20CR. Apart from deviations on the decadal timescale, the correlation-map between the wind index and HiResAFF (Fig. 3-9) shows a reasonable annual co-variability over wide areas of the North Sea and southern NE-Atlantic, central Europe and the Baltic Sea region south of 60°N. The highest correlations of up to 0.5 exist over the British Isles and Northern Germany to Poland.

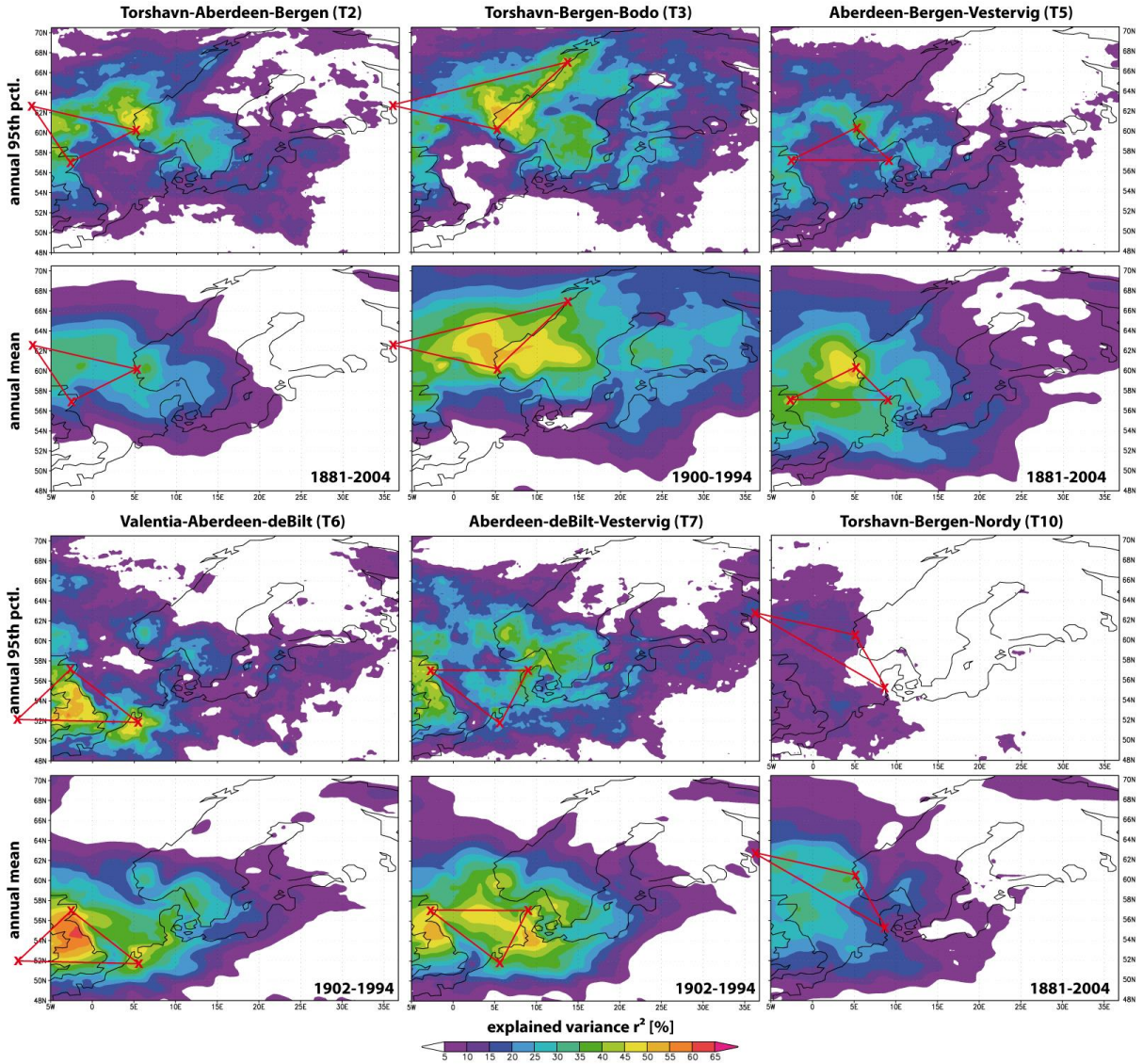

Fig. 3-7: Explained variance between annual time series of geostrophic wind speeds from different station triplets (Krueger et al., 2013) and fields of HiResAFF for annual mean and annual 95th percentile wind speeds. Note different time periods covered by the triangles.

Baltic Sea region – Finland since 1884

The two longest geostrophic wind indices by Suvilampi (2009) derived from station triplets over southern (T1, 1884–2007) and central Finland (T3, 1888–2007) (Fig. 3-8) are compared with time series of annual 95th percentile wind speeds of HiResAFF and the ensemble mean of 20CR. While the geostrophic indices T1 and T3 share only a correlation of r=0.43 (r=0.36 detrended), the corresponding time series of HiResAFF and 20CR are almost identical and only T1 is presented. While geostrophic wind indices in Fig. 3-8 show a monotonic decrease from the 1890s to 1970s, 20CR and HiResAFF do not show such a long-term change. Low-frequency variations of 20CR agree well with HiResAFF with exception of the 1970s being again not unusually calm in 20CR. Both geostrophic wind indices show a strong deviation in the 1980s which is not supported by other time series.

The correlations of annual time series with HiResAFF and 20CR are relatively low with r=0.26 (0.31 detrended) and r=0.35 for both datasets with T1, respectively. With T3, correlations with 20CR drop to r=0.18 and to r=0.17 (0.27 detrended) for HiResAFF. No

agreement exists regarding long-term trends with weak significant positive trends of +0.18 m/s (+2.3%) per century in HiResAFF (T1), no trend in 20CR and significant negative trends for T1 with -0.56 m/s (-2.81%) and -1.25 m/s (-6.3%) per century for T3. A one-point-correlation-map of the T1 index with HiResAFF shows low but statistically significant ($p<0.05$) correlations on the annual scale in remote regions north and west of Finland. This result is almost identical with the T3 index which itself shows little to no significant correlation within the area of T3.

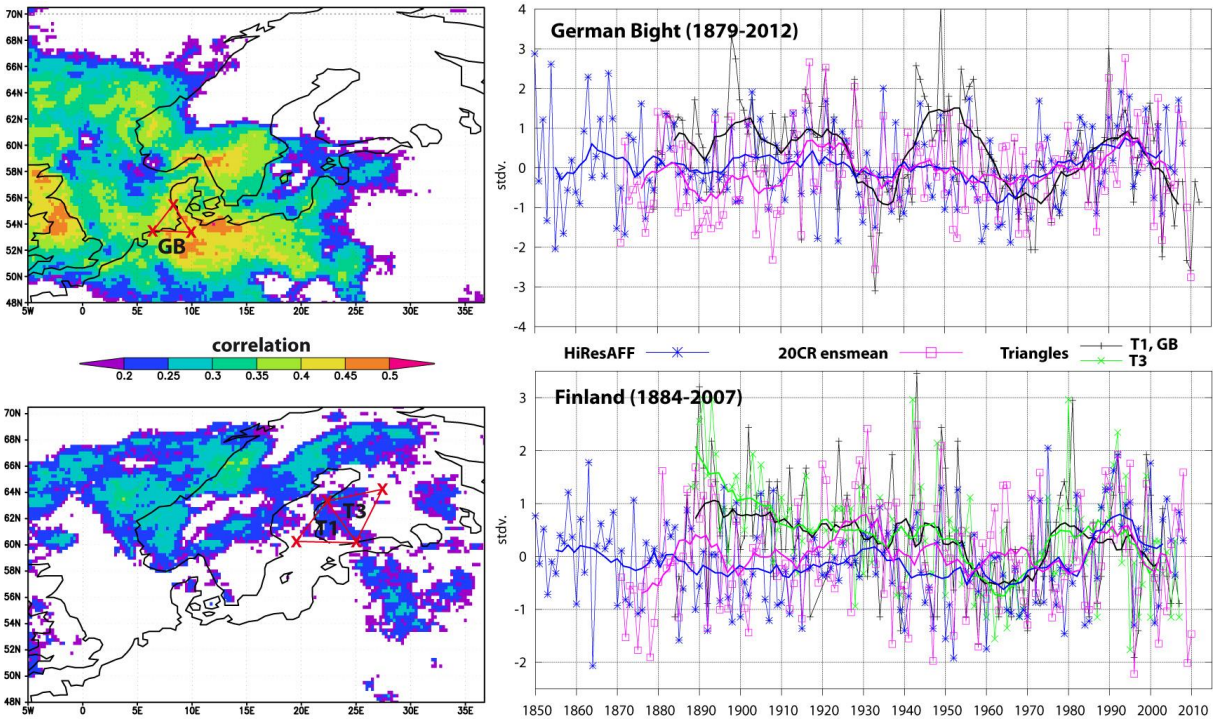

Fig. 3-8: Comparison of time series of annual storminess derived from 95th annual percentiles daily geostrophic wind speeds over the German Bight 1879-2012 (GB, updated after Schmidt and von Storch, 1993) and Finland 1884-2007 (T1, T3, Suvilampi, 2009) with 95th annual percentile wind speeds of HiResAFF. Additionally shown is the corresponding 20CR time series. Correlation maps of geostrophic indices with HiResAFF (white: not significant at p<0.05).

Baltic Sea region – Stockholm since 1850 (1785)

Annual and decadal running mean time series of high annual wind percentiles from HiResAFF since 1850 and 20CR since 1871 are compared with the longest wind indices over Baltic Sea region derived from pressure records at Stockholm. The single-station proxies of the annual number of deep lows (N<980 hPa), the 99th percentile of pressure changes within eight hours (99th pctl. dp/8h) and the annual number of days exceeding 25 hPa pressure changes (dp>25hPa/24h) are taken from Bärring and Fortuniak (2009).

As shown in Fig. 3-9, a relative good agreement exists for the low-frequency variations of wind indices and the annual 99th percentile wind speeds of HiResAFF with stormy periods in the 1870s and 1990s and an unusually calm period in the 1970s. Only the number of deep lows does not reflect the calm period. 20CR agrees only partly with the different indices and the reconstruction and shows remarkably stormy up-ticks in the late 1950s and 1930s and very low decadal values before around 1890. Apart from 20CR, a typical feature of all time series is a rather monotonic decrease from the stormy 1870s to the calm 1970s.

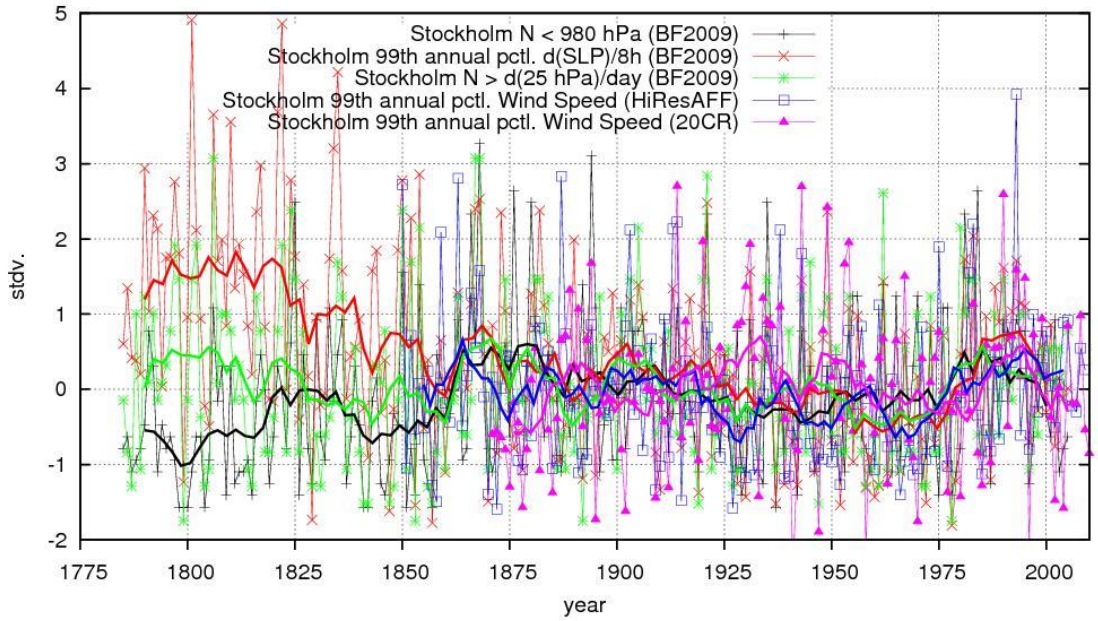

Fig. 3-9: Storm indices of the annual number of deep lows (N<980 hPa), the 99th annual percentile of pressure tendencies per 8h, the annual number of days exceeding a pressure tendency of 25 hPa derived from the station of Stockholm 1785-2005 (Bärring and Fortuniak 2009) compared to the reconstructed annual 99th percentile of wind speeds in the vicinity of Stockholm 1850-2008 from HiResAFF and 20CR from 1871-2008. Time series are normalized with respect to 1958-2005. 11y-running means are used to highlight decadal variations.

The large increase of the pressure changes within eight hours and the concurrent decreasing number of deep lows before around 1850 are caused by timely irregular pressure observations (Bärring and von Storch, 2004) which is a less serious problem in case of daily pressure changes (green line). To quantify the co-variability between HiResAFF and the storm indices, rank correlations are calculated for the period 1850–2005. As reflected by time series in Fig. 3-9, the poorest agreement with HiResAFF is found for the annual number of deep lows with only r=0.22 followed by the annual number of days exceeding 25 hPa pressure changes per day with r=0.38. The best agreement is found for the annual 99th percentile of pressure changes per eight hours with r=0.50. None of the time series point to notable long-term changes after 1850 and 1871, respectively.

As pressure changes at a specific site are not necessarily linked to storminess over the respective area, one-point-correlation maps are calculated for all three indices with HiResAFF (Fig. 3-10). Rank correlations are used because they provide a better basis for comparisons between the different indices. For annual time series, the highest correlations of around r~0.5 are generally located west and south of Stockholm over Northern Germany (N<980 hPa, N>25 hPa/day) and around the southern and central Baltic Sea region (99th pctl. dSLP/8h). North of Stockholm, the N<980 index does not show any significant skill with little skill also for N>25hPa/day. The index of annual 99th percentile pressure changes per eight hours shows the highest and spatially most extended skill. Very low skill exists along for near-coastal storm variations along eastern UK.

On decadal time scales, the spatial patterns are relatively similar, with lowest skill for deep lows and very high correlations for extreme pressure changes within eight hours. Only over the North Sea, the latter index seems to be not very well suited to predict decadal storm variations. Significance levels are again estimated through block-bootstrapping to account for the high serial correlations of time series, as noted previously.

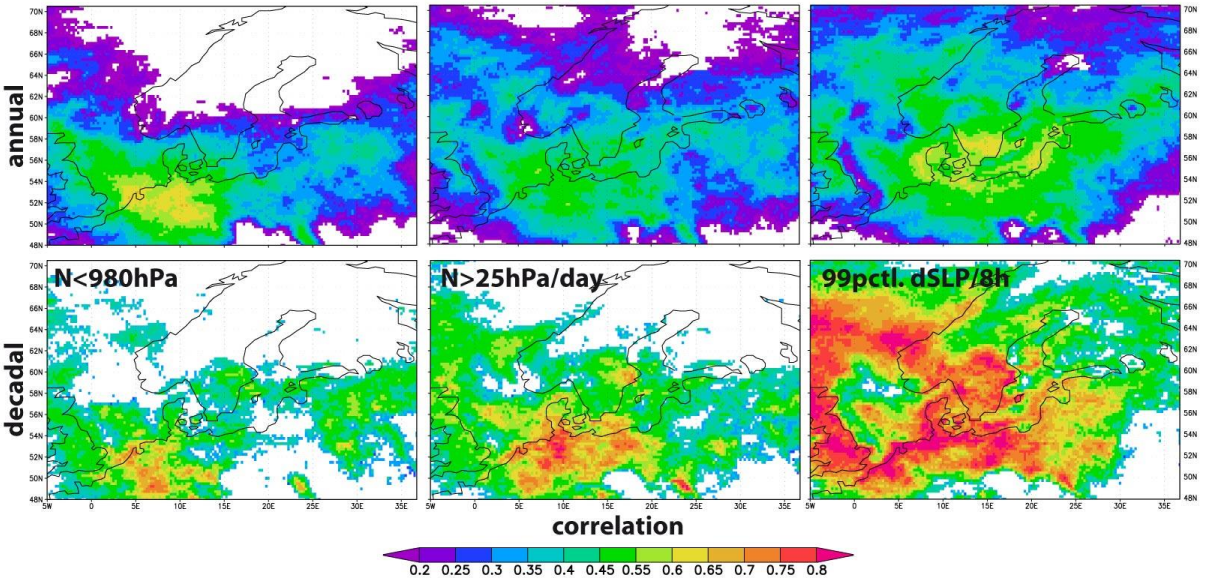

Fig. 3-10: One-point-correlation maps of different storm indices of Stockholm with annual 95th percentile wind fields of HiResAFF 1850-2008. White: non-significant correlation (p<0.05).

3.3.6 Long-term trends in other regional storm studies

In order to get a more conclusive view on the potential existence of regional long-term trends over the Euro-Atlantic region, Fig. 3-11 presents a synthesized literature review about historical storm or storm related observations. The keywords *wind, storm or storminess* are used to find relevant studies published since 1990 in the online database of *Science Direct* (www.sciencedirect.com) followed by a snowball search. The published records are grouped into three categories regarding long-term trends in wind, storms or wind and storm related proxies. The different studies, the time period and the result regarding the existence of long-term trends are summarized in Fig. 3-11 for different sub-regions.

Many studies do not explicitly examine trends and/or do not test the robustness of trends by applying significance tests on resulting trends. To also evaluate the robustness of different trend estimates in Fig. 3-11, robust results are indicated with dark/solid colours with dark red and dark blue colours indicating significant positive/negative trends. Dark green indicates no significant trend. In other cases, a subjective category is assigned to the visible tendency of the records and the rather descriptive results are indicated by light red, light blue and light green colours. For the last four to six decades, only some examples of different storm studies are given.

The results for the different studies on historical wind and storm observations show a heterogeneous picture depending on the timescale and region and also on the proxy. Taking the trends at least since around 1900 or further back in time, the NE-Atlantic region including Iceland shows no robust long-term trends with the exception of a single-station proxy for SW-Iceland for 1823–2006 (Hanna et al., 2008) and a weather type classification over the Euro-Atlantic region 1881–1992 (Schiesser et al., 1997). Both results show however a high level of uncertainty. The first study uses the annual mean of absolute pressure changes per 24h from single-station pressure, which already yields no significant trends over the southern and eastern parts of Iceland since 1833. The second study shows a significant positive trend in the frequency of west cyclonic weather types based on historical weather maps. The detection of such cyclones might be less reliable back in time so that this trend might be spurious.

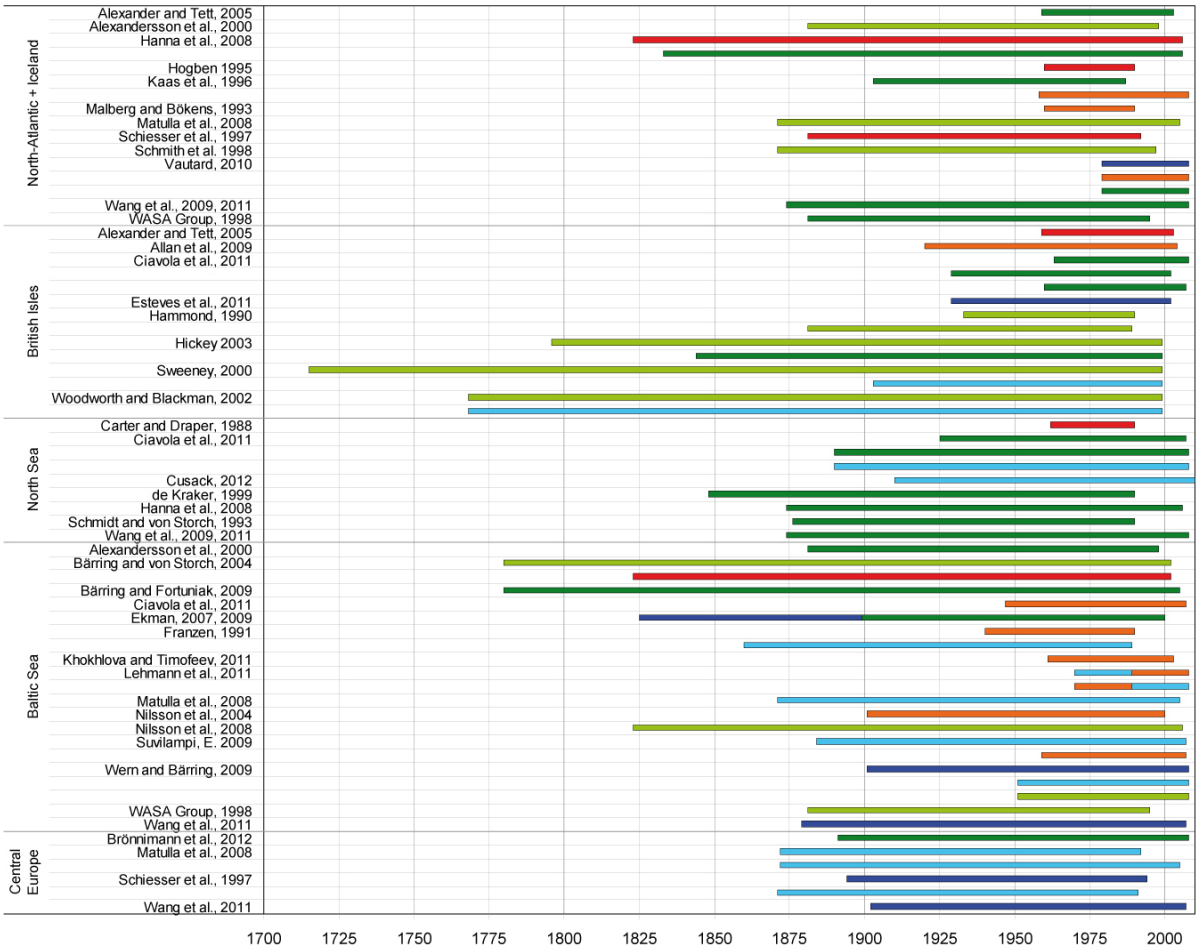

Fig. 3-11: Synthesis of publications on regional wind and storminess trends estimated from direct or indirect historical records on the wind climate over the Euro-Atlantic region. Dark colours of red, blue and green indicate positive, negative or non-significant trends (at least $p<0.05$). Light colours are used where no objective trend analysis was possible/applied by the respective authors.

Numerous studies encompassing a wide range of different sources agree in terms of the absence of any long-term trends since 1715 or later over the British Isles and North Sea. If any, annual maximum surges at high water in Liverpool since 1768 or surge information as a proxy for storm frequency for the Netherlands since 1890 (Ciavola et al., 2011) indicate a negative tendency for storminess in this region. The sources reach here from documentary storm reports since 1715 in the Dublin region (Sweeney, 2000) over surge-level records at Liverpool (Woodworth and Blackmann, 2002) or historical wind records related to the annual number of gale days since 1883 at Armagh (Ireland) (Hickey, 2003) as well as pressure-based geostrophic wind indices since the 1880s (e.g. Wang et al., 2009a, 2011).

Over the Baltic Sea region, the different studies including those discussed in chapter 3.3.5 mostly suggest no robust long-term trends or even a negative tendency since 1900 or earlier. Different direct wind observations at the Northern Alpine region of Switzerland (Schiesser et al., 1997) and Austria (Matulla et al., 2007) as well as geostrophic storm indices over central Europe (Matulla et al., 2007; Wang et al., 2011) indicate a clear negative tendency. This is mainly caused by a relatively weak storm period in the 1990s compared to other regions while the 1880s have been similarly stormy as over the North Atlantic. An extended discussion of studies in Fig. 3-11 are published in upcoming climate assessment reports (BACC-II, NOSCCA) and a review paper on storminess (Feser et al., 2014).

3.3.7 Inconsistencies in the 20th Century Reanalysis

The previous analysis of decadal storm variations (Fig. 3-4) and centennial trends (Fig. 3-6) indicate large deviations between HiResAFF and 20CR. The strong increase in 20CR's storminess is obviously also inconsistent with other historical wind information over the Euro-Atlantic region (Fig. 3-11). Figure 3-12 shows that decadal correlations between annual 95th percentile wind speeds of 20CR and HiResAFF show a clear improvement in some regions when the trend has been removed before (Fig. 3-12b). The difference between the original and detrended case clearly identifies the Norwegian Sea and North Sea as region of main discrepancies (Fig. 3-12c). The improvement in the detrended case suggest that both datasets would show a good co-variability without the spurious long-term trends in 20CR.

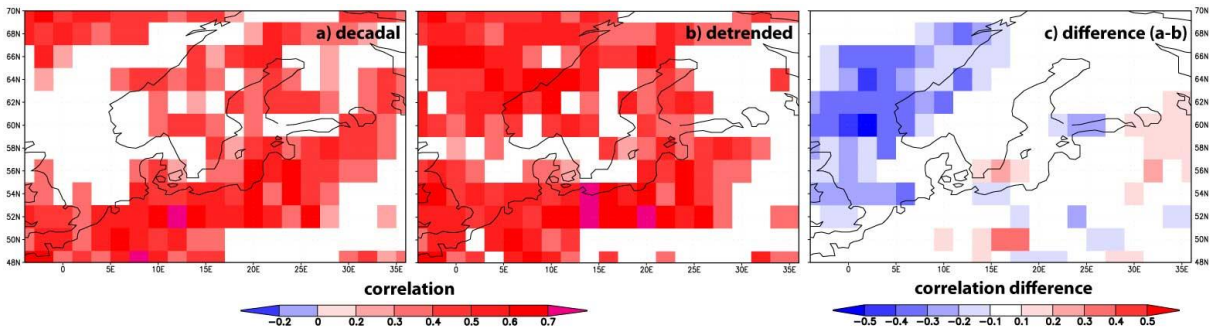

Fig. 3-12: Correlation maps of 11yr running means of annual 95th percentile wind speeds of HiResAFF and 20CR. a) original time series, b) after previously detrending the data and c) difference of correlations of original minus detrended results.

Krueger et al. (2013) find even larger deviations in low-frequency variations further west over the North Atlantic. In this study, the averaged and filtered time series of ten geostrophic wind indices calculated from pressure observations have been compared with similarly calculated indices from closest grid points of 20CR. The correlation of filtered time series before 1940 was found to be only 0.11 but 0.95 afterwards. Using the unfiltered annual time series of this study, the removal of long-term trends from 20CR and the geostrophic index yields also a clear improvement for the annual correlations over the period 1881-2004 or 1881-1939 while no improvement is achieved for the period 1940-2004 (Fig. 3-13). Taking the grid point of the highest correlation of the averaged 56 correlation maps in Fig. 3-13, the averaged explained variance is 20.09% for the period 1881–2004 and doubles to 41.03% in the detrended case.

The very similar correlations of the original vs. detrended data between HiResAFF and geostrophic indices used by Krueger et al. (2013) in Fig. 3-7 confirms that the inconsistencies in low-frequency variations only exist in case of 20CR. A closer look on running correlations for different sub-periods between HiResAFF and 20CR (not shown) indicates large mostly abrupt changes for different regions over time, eventually pointing towards spatiotemporal inconsistencies in 20CR. A short preliminary analysis of a 31yr running EOF analysis shows that the variance represented by the first leading mode in 20CR decreases back in time and hence the "noise" by other modes increases compared to HiResAFF. After 1874, HiResAFF makes use of a relatively stable number of predictors and the distance between analogs on basis of the 23 stations does not indicate profound changes. A more thorough analysis is needed to estimate 20CR's robustness over time e.g. by considering changes in running correlations with HiResAFF or other long-term storm proxies in the context of variations in the number of assimilated stations into 20CR. This is however beyond the scope of this study.

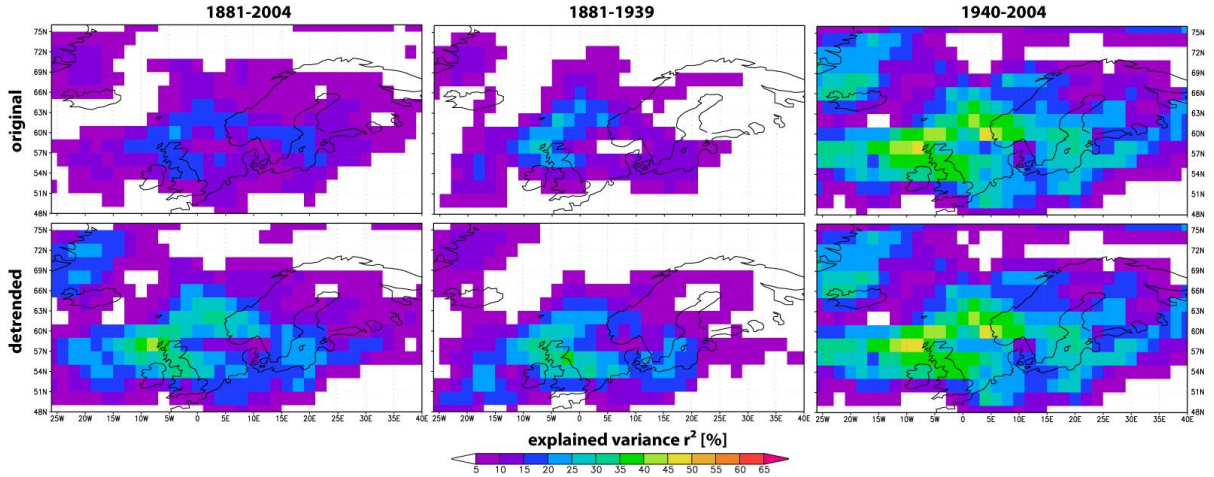

Fig. 3-13: One-point-correlation maps of annual 95th percentile wind fields of 20CR and the geostrophic wind index from pressure observations (updated from Alexandersson et al., 2000) for different time periods with and without previously detrending the data. Results present the averaged explained variance from all 56 ensemble members.

3.4 Discussion

3.4.1 Consistency of long-term variations and trends

The co-variability of the first leading mode (PC1, Fig. 3-2) of annual storm variability over Northern Europe between HiResAFF and 20CR is mostly high on annual to decadal scale. However, clear systematic deviations in the level of storminess exist. In agreement with other pressure-based storm indices over the NE-Atlantic, central Europe to southern Scandinavia (e.g. Alexandersson et al., 2000; Matulla et al., 2007; Bärring and Fortuniak, 2009; Wang et al., 2009a; 2011), HiResAFF shows a rather monotonic decrease of storminess from the stormy period around 1890 to the very calm 1970s. 20CR does not show such a monotonous decrease as the period before 1900 shows very low storm activity while the calm 1970s in observations remain on rather average conditions in 20CR. This leads to a significant positive long-term trend in 20CR (Donat et al., 2011) which is obviously inconsistent with HiResAFF and other observational storm indices (Krueger et al., 2013). Also storminess related to deep lows (PC2, Fig. 3-2) shows a high co-variability on annual to decadal scale with systematic discrepancies for two periods. Again, 20CR does not reflect the very calm 1970s and shows in addition much lower storm levels during the 1910s. This leads again to a significant positive trend of 20CR's PC2 while no trend exists in case of HiResAFF since 1871.

Although the deviations in PC1 between HiResAFF and 20CR mirror the same deviations found for geostrophic storm indices over the NE-Atlantic since 1881, the discrepancies are however smaller than reported earlier by Krueger et al. (2013) over the North Atlantic region. One explanation for the good agreement between HiResAFF and 20CR for the first leading modes of storm variability can be related to the EOF analysis itself. Considering only the first two leading modes, any noise contained in EOFs with a higher order but low explained variance (<5%) is disregarded by focusing only on the co-variability of the leading co-variance matrices of both datasets. The analysis of the large-scale co-variability is hence insensitive to smaller scale spatial discrepancies. In addition, the assumption of a spatiotemporally stationary pattern for different EOFs does not account for potential changes over time. A 31yr running EOF analysis partly indicates increasing differences in the amount of variances explained by the first leading modes between 20CR and HiResAFF (not shown). Relative to the reconstruction, the leading mode of 20CR’s storminess explains increasingly

less of the total variance back in time. As the number of predictors and the distance between analogs remains relatively constant since 1874 in case of HiResAFF, the increasing "noise" in 20CR back in time could be related to the decreasing number of stations assimilated into the model in the early period – and hence to potential regional discrepancies.

Inconsistencies mainly over remote sea areas

Spatiotemporal differences in the number of assimilated stations by 20CR and a different regional skill of HiResAFF (Schenk and Zorita, 2012; chapter 2) lead to differences in the agreement of regional storm variations. The comparison of annual storm variations between 20CR and the reconstruction mostly yields reasonable to high correlations with non-significant correlations over the SE and NW corner of the domain (Fig. 3-5). The latter may be attributed at least partly to the reconstruction as the skill for monthly mean wind speeds was already low in these regions for the validation of HiResAFF in the period 1958-2007 (Fig. 2-6). The explanation for a reduced confidence into the reconstruction in these regions is low availability of predictors over the NE-Atlantic and the more continental wind climate towards the SE, where wind variations are less strongly linked to large-scale pressure variations predicted through SLP (Schenk and Zorita, 2012, chapter 2). The reduced agreement over sea areas rather points to a lower skill of 20CR as the validation of HiResAFF showed here very high correlations on monthly scale. Also high correlations with geostrophic indices over the UK, North Sea and NE-Atlantic from Alexandersson et al. (2000) confirm a realistic reconstruction of HiResAFF over sea areas (Fig. 3-7).

To define the region and timing when low-frequency deviations relative to 20CR occur, decadal running mean time series of annual 95th percentile wind speeds (Fig. 3-4) are presented for both datasets for different regions. Consistent with the findings of Krueger et al. (2013), clear regional differences start to increase before around 1940. However, deviations on decadal scale are partly less pronounced over land areas after the 1930s compared to open sea areas. As a result, no significant decadal scale correlations between both datasets exist mainly over open sea areas where the time series show opposing trends before 1920 over the NE-Atlantic and North Sea (Fig. 3-5). Correlations of detrended data of 20CR relative to HiResAFF (Fig. 3-12) or geostrophic indices over the North Atlantic (Fig. 3-13) lead to clear improvements over sea areas. This confirms that all datasets show mostly good annual co-variability if inconsistent trends are removed from 20CR. The small internal model spread of trends between the 56 ensemble members (Fig. 3-6) shows that this result is valid for all members of 20CR. In contrast, no improvement is achieved by detrending when the geostrophic indices are correlated with HiResAFF. This implies that the updated geostrophic storm indices used by Krueger et al. (2013) and the reconstruction of HiResAFF provide a consistent and robust representation of annual and low-frequency storminess for at least 120 years.

As the number of predictors decreases before 1874 (Tab. 3-1) in case of HiResAFF, a test reconstruction based on only six stations around the central domain was used to test potential deviations of low-frequency variations and trends in remote regions to the predictors. The evaluation of Test_N1850 shows indeed that low-frequency variations (Fig. 3-4) and trends (Fig. 3-6) over distant regions of the NE-Atlantic deviate from HiResAFF while little difference exists over the rest of the domain. As a consequence, the early period over the NE-Atlantic should be taken with care while other regions provide a robust reconstruction back to 1850. The sensitivity study with a very low but constant number of predictors hence confirms that a restriction to a relatively fixed number of continuously available predictors as in case of HiResAFF should be maintained rather than including all available stations over time.

Different levels of agreement with pressure-based storm indices

A very high agreement of annual to centennial mean wind and storm variations exists between HiResAFF and different pressure-based geostrophic mean annual and annual 95th percentile wind indices over the North Sea and NE-Atlantic since 1881 (Fig. 3-7). The validation is however not fully independent as at least some stations used by Alexandersson et al. (2000) are also used as predictors of HiResAFF (Tab. 3-1). Nevertheless, the very high correlations confirm a highly realistic reconstruction over these sea areas where only one station (Torshavn) was available as predictor for the western domain of the reconstruction.

The comparison of annual storm variations of HiResAFF with the geostrophic storm index over the German Bight since 1879 shows lower correlations on annual scale and little agreement for decadal scale variations (Fig. 3-8). As the pressure data originally used by Schmidt and von Storch (1993) has not been homogenized so far (Rosenhagen, pers. com.), the very high storm levels of this index during the 1950s to 1960s seem to indicate an offset which coincides with the usage of secondary stations during this period (Norderney instead of Borkum in 1939 and 1947-49, Fuhlsbüttel instead of Hamburg 1938-50, 1954-55, see Schmidt and von Storch, 1993). Also 20CR and high annual water level percentiles for the German Bight recorded over the last 109 years do not show increased storminess during this period (Dangendorf et al., 2013).

Also the geostrophic indices over Finland since 1884 (Suvilampi, 2009) share only a low correlation with HiResAFF. Here, a clearly higher agreement is found between HiResAFF and 20CR on annual to decadal scales. Sharing two common stations, the very low agreement of T1 and T3 indices indicate a low confidence into the used pressure data over Finland. Another reason for the low annual correlation might be related to the rather continental wind regime over Finland where stronger ageostrophic effects weaken the link to large-scale geostrophic wind variations.

The comparison of annual storminess of HiResAFF since 1850 with different pressure-based single-station storm indices of Stockholm (Bärring and Fortuniak, 2009) does not provide an independent validation as Stockholm is also used as a predictor. The high agreement on decadal scale confirms however that analog-upscaling yields very similar low-frequency variations as single-station indices. As noted already in a previous study of these indices (Bärring and von Storch, 2004), the different indices describe different properties related to storminess. Consistent with the model-based evaluation of single-station storm proxies by Krueger and von Storch (2011), annual correlations with HiResAFF's real (model) wind speeds are rather low for deep lows but reasonably high for strong pressure changes per time unit. One-point-correlation maps with these indices show generally a low skill at the station itself and mostly non-significant correlations northward. The otherwise good to very high annual and decadal correlations southwest to the station simply refers to the fact that a deep low or strong pressure change per time unit observed at Stockholm detects mostly the centre of a passing low while the storm field of these meso-scale lows lays traditionally SW to W of the centre following the cold front. In part, very low correlations along the eastern coast of the UK indicate that the reconstruction reproduces also regional topographically influenced wind regimes which cannot be represented by the large-scale geostrophic storm indices.

The absence of robust long-term trends in HiResAFF and pressure-based storm indices since the 1890s or earlier are in good agreement with other historical long-term observations in the scientific literature (Fig. 3-11). Some indices point to rather weak negative trends and none of them support the strong significant positive trends of 20CR (Donat et al., 2011).

3.4.2 Analysis of large-scale storm variations since 1850

Based on an EOF analysis of annual storminess, the first two leading modes of annual storm variability over Northern Europe represent roughly 50% of the total variance. The regression of PC1 with annual 5th percentile pressure fields shows that about one third of the storm variability relates to variations in the zonal pressure gradient while about 15% can be explained by variations related to the strength of deep lows over Scandinavia (PC2). As storm activity over Northern Europe is strongly linked to variations of zonal pressure gradients over the North Atlantic (NAO), the PC1 time series of HiResAFF clearly reflects the time evolution of other long-term annul storm indices over the NE-Atlantic (e.g. Alexandersson et al. (2000); Matulla et al., 2007; Wang et al., 2009a, 2011) or Stockholm (Bärring and von Storch, 2004; Bärring and Fortuniak, 2009, Fig. 3-10). Similar to HiResAFF, these indices reveal very stormy periods in the 1870s, 1890s and 1990s and a very calm period in the 1970s. Consistent with these studies, HiResAFF does not show any significant trend in annual storminess since 1871 or 1850 over Northern Europe.

The fraction of annual storm variations related to variations in the strength of deep lows over Scandinavia of PC2 indicates also no long-term change for HiResAFF since 1850 or 1871 but a significant increase in 20CR since 1871. The positive trend for 20CR's PC2 results here from clearly higher anomalies during the 1990s compared to HiResAFF while 20CR does not capture the very low anomalies during the 1970s. The time series of the annual number of deep lows in HiResAFF suggest also no long-term change over Northern Europe since 1850. The very stormy/calm periods mentioned before are also clearly reflected by very high/low numbers of deep lows. Different to the single-station count of annual N<980 hPa at Stockholm (Bärring and Fortuniak, 2009, Fig. 3-9) and PC2 of 20CR (Fig. 3-2), HiResAFF shows a clear minimum in the 1970s also for the number of deep lows.

Indications for spatial changes in the area affected by deep lows and storminess

Apart from potential long-term trends in storminess, numerous studies for the period of the last 40-60 years find clear evidence for increased storm activity north of around 55-60°N over the North Atlantic together with a NE shift of storm tracks and deep lows towards Scandinavia since the 1970s (Sickmöller et al., 1997; Chang and Fu, 2002; Wang et al., 2006, Lehmann et al., 2011). This together with decreasing storminess southwards (Gulev et al., 2001; McCabe et al., 2001; Trigo et al., 2006; Raible et al., 2008) appears to be consistent with scenario simulations for 2100 under increased greenhouse gas concentrations (e.g. Ulbrich et al., 2009). For the North Atlantic, Wang et al. (2009b) find that external forcing played a role in the wind and wave climate since 1950 but less likely before since 1900.

A strong limitation of all these studies relate to the short time scale being mostly limited to the last 40-60 years. Even the detection and attribution study by Wang et al. (2009b) since 1900 might be too short given the large multi-decadal variations in extra-tropical storm variations over the Euro-Atlantic region. Besides not showing any long-term trends since the 16th century (Fig. 3-11), most of these studies highlight the large (multi-)decadal variations in the storm climate over several centuries comparable to those seen in the 20th century. However, a generally negative long-term trend seems to exist over central Europe and along the Northern Alps since around 1870 (Schiesser et al., 1997, Matulla et al., 2007, Wang et al., 2011). This can be explained by the NE-shift of storm tracks mentioned earlier since the 1970s. Here, storms were directed less frequently towards southern regions during the 1990s (Schiesser et al., 1997) while the end of the 19th century has been similar stormy over all regions.

Analysing the running decadal mean area affected by deep lows over Northern Europe since 1850, HiResAFF suggests that the NE-shift and increasing number of deep lows from the 1970s to 2000 is unprecedented in the last 160 years (Fig. 3-3). This implies that the area of the equally stormy period in the 1870s has faced only a modest impact of deep lows. Although the last two decades of the 20th century show a spatially very large influence of deep lows, the very low values in the 1970s and low values from the 1920s to 1980s lead to no significant long-term increase. Whether external forcing contributes to the unprecedented 1990s or whether several decades just reflect large natural variations needs to be further studied.

The principal difference between the storm periods of the 1870s and 1990s can be seen in very cold conditions during the first period while the second storm episode falls within the warmest decades since at least 1850 (chapter 4). If the warmer storm period favoured larger deep lows as a result of higher SST's over the North Atlantic than at the end of the Little Ice Age around 1850-70 or whether this is mere coincidence should be subject to further studies. At least externally forced model studies with coupled AOGCMs for the last millennium do not suggest any robust link between temperature and storm variations (Fischer-Bruns et al., 2005) or cyclone activity (Xia et al., 2012) over the northern hemisphere. This view is supported by HiResAFF and a wealth of different historical storm studies for the Euro-Atlantic region of the last three centuries summarized in Fig. 3-11. Only the NE-shift of storm tracks since the 1970s, the resulting negative storminess trend over central Europe together with the unprecedented high spatial impact of deep lows at the end of the 20th century (HiResAFF and Lehmann et al., 2011) render the 1990s a different storm period compared to the 1870s while storminess level kept within the range of centennial variations.

3.5 Conclusion

The reconstruction of HiResAFF (Schenk and Zorita, 2012; chapter 2) has been extended to the period 1850-2009 based on up to 23 stations providing daily SLP observations. The comparison of annual storm variations of HiResAFF with different pressure-based storm indices shows a good agreement with geostrophic storm indices over the NE-Atlantic and North Sea since 1881 and single-station proxies derived for Stockholm since 1850. For geostrophic storm indices over the German Bight and Finland, the comparison points towards inhomogeneity in the currently not homogenized pressure data of these indices. In the latter cases, a clearly higher agreement is found here between 20CR and HiResAFF i.e. on decadal scale with exception of the period before around 1920.

The comparison of 20CR and HiResAFF regarding annual storm variations since 1871 shows a relatively good agreement on inter-annual scale with 80% of grid points showing statistically significant correlations at the 5% level (Fig. 3-5a). Little improvement is achieved when trends are removed prior to the correlation. Clear inconsistencies exist however on decadal scale mainly over open sea areas. These are greatly reduced when the long-term trends are removed before correlating the decadal co-variability of both fields (Fig. 3-13c). The consistency of HiResAFF with pressure-based storm indices and the clearly improved correlations of a detrended 20CR with these indices and HiResAFF confirms a good long-term skill of storm indices, HiResAFF and the detrended 20CR. Together with the results of Krueger et al. (2013), HiResAFF clearly attributes the inconsistencies of low-frequency variations and trends to be caused by the 20CR data assimilation strategy in the early period prior to around 1940 to 1920. Different to the study by Krueger et al. (2013) located over the North Atlantic, HiResAFF suggests partly less pronounced inconsistencies in 20CR over land areas. The comparison of HiResAFF with 20CR hence confirms the two so far existing

studies for long-term storminess in 20CR. The strong discrepancies over the sea areas of the NE-Atlantic are in line with the results by Krueger et al. (2013) while the better agreement over land areas supports the study by Brönnimann et al. (2012) for the northern Alps. As pressure observation are hardly available over sea areas back in time, the strong discrepancies over these areas clearly support the hypothesis that strong changes in the amount of assimilated stations introduce spurious trends into reanalysis datasets over data sparse regions. The test reconstruction of HiResAFF with only six stations supports the robustness of the original reconstruction with exception of the remote region over the NE-Atlantic before around 1870. The sensitivity study Test_N1850 suggests that analog-reconstructions should try to keep the number of predictors constant rather than including all available information over time.

Regarding the analysis of storminess, HiResAFF together with homogeneous pressure-based storm proxies show that storminess during the 1990s has been at similar levels as in the 1890s or 1870s with a very calm period in the 1970s. With exception of the inconsistent positive trends in 20CR, no robust trends are found for the last 130 to 160 years and even a slight negative tendency is supported by some historical storm proxies since the 1880s.

The extended analysis of the annual number of deep lows derived from HiResAFF since 1850 over Northern Europe yields an unprecedented high spatial extension affected by deep lows at the end of the 20th century. This can be seen as a regional imprint of the NE-shift of deep lows over the North Atlantic towards Scandinavia observed in reanalysis data since the 1970s. This shift explains also less frequent storms over central Europe and the northern Alps during the 1990s which leads to a negative trend in this regions (Fig. 3-11). The results derived from HiResAFF since 1850 and different studies on the historical storm climate over the Euro-Atlantic region of the last centuries lead to a clear conclusion:

(1) Storm levels in the 1990s have been on similar levels as in the 1870s or 1890s over Northern Europe.

(2) A NE-shift of the extra-tropical storm track and hence deep lows towards Northern Europe is unprecedented since at least 1850 and is also reflected by negative trends in storm frequencies over central Europe and northern Alps since 1870.

(3) HiResAFF and storm indices based on homogenized pressure data show consistent annual and decadal scale variations over the last 120-160 years while 20CR is subject to strong inhomogeneity before around 1920 to 1940 with spurious long-term trends mainly over open sea areas.

4 Reconstruction of long-term variations and trends of temperature fields since 1850

The reconstruction of temperature fields through statistical analog-upscaling in chapter 2 has been shown to realistically reproduce daily and monthly correlation and variance in the validation period 1961-2007. In this chapter, the extended reconstruction for the period 1850-2009 by the same method and predictors is evaluated. It is tested whether the available analogs from the period 1961-2007 allow a realistic reconstruction of long-term variations, trends and spatial patterns of seasonal temperatures compared to observations since 1850. Besides an evaluation of the reconstruction's robustness over time, the dependency on deficiencies introduced by the used model fields relative to observations are assessed through a cross comparison of HiResAFF, model fields of RCAX and gridded observations.

4.1 Introduction

Northern Europe and the Baltic Sea region have faced strong warming trends over the last 140-160 years on annual and seasonal basis partly with exception of summer (Jones and Moberg, 2003; BACC, 2008). Based on gridded temperature observations of CRUTEM3 (Brohan et al., 2006), updated annual warming trends over the Baltic Sea region have been strongest north of 60° N with +1.1 K per century over the period 1871-2011. This is clearly stronger than the warming trend over the entire northern hemisphere with +0.76 K per century (+1.14 K) over the period 1861-2010 (Jones et al., 2012) and comparable to the southern Baltic Sea area with warming trends of +0.8 K.

Besides the centennial warming trend, temperatures over Northern Europe are characterized by large inter-annual to (multi-)decadal variations. In the winter half year, these variations are to a large extent explained by variations of the North Atlantic Oscillation (NAO) (Hurrell, 1995; Hurrell et al., 2003). In addition to strong positive or negative NAO phases, large seasonal temperature anomalies are also related to variations in the frequency of large-scale atmospheric blocking over the Euro-Atlantic region. Increased blocking frequencies during winter contributed to the cold period of the 1960-70s and the low blocking activity during the 1990s are associated with very warm and humid winters (cf. Rimbu and Lohmann, 2011). Besides these large decadal variations in the frequency and also location of atmospheric blocking, an overall significant negative trend for the number of blocked days exists over the Euro-Atlantic region in the period 1955-2004 (Luo and Wan, 2005) or 1948-2002 (Barriopedro et al., 2006) based on NCEP-reanalysis data.

Although Northern Europe is dominated by large temperature variations, significant warming trends can be clearly detected and partly attributed to anthropogenic forcing (Bhend and von Storch, 2009). Due to the semi-enclosed and mostly shallow basin of the Baltic Sea, warming trends are expected to contribute to the large changes in environmental conditions of the Baltic Sea ecosystem since the end of "Little Ice Age" (LIA) around 1877 (Omstedt and Chen, 2001) together with anthropogenic eutrophication i.e. since the 1950s (chapter 5, Meier et al., 2012b, Gustafsson et al., 2012). As only very little bio(geo)chemical and physical measurements exist to capture long-term changes in ecosystems, ecosystem models driven by high resolution meteorological fields like HiResAFF (Schenk and Zorita, 2012) are required on long timescales at least since 1850 to 1870. In the case of temperature, the forcing fields should provide realistic seasonal long-term variations and trends with physically consistent temperature conditions over sea areas where little or no observations exist back in time.

Existing gridded temperature fields before 1900 are not very well suited to provide the necessary forcing conditions for marine ecosystem models due to their very coarse resolution. The most recent fourth version of CRUTEM4 (Jones et al., 2012) provides monthly mean temperatures with a horizontal resolution of only 5°x5° (~500 km) since 1850 with missing data till around the 1880s for most northern and eastern parts of Fenno-Scandinavia. These fields do not consider regional details and terrestrial temperatures are interpolated over the entire Baltic Sea area. The reconstructed temperature fields of HiResAFF are intended to provide an alternative dataset which shows the same low-frequency variations and trends but with higher spatial resolution and physical consistency. The advantage of analog-upscaling instead of gridding and interpolation is here that the necessary spatial temperature patterns including sea areas are provided by taking fields from a regional climate simulation while the temporal variations of these fields are predicted by monthly station observations.

In the following, the ability of the analog-method is evaluated regarding a realistic reconstruction of long-term variations and trends of seasonal temperatures and their corresponding spatial patterns of seasonal trends. In a first step, a cross-comparison over the period 1961-2007 is applied for HiResAFF, the regional climate simulation RCAX (chapter 2) and gridded temperature observations of CRU TS 3.10 (Harris et al., 2013). This allows to identify to which extent the reconstruction skill is dependent on the used model fields of RCAX. In a second step, the comparison with CRU TS 3.10 is extended to the period 1901-2009 for seasonal trends and their spatial patterns before the long-term warming since 1850 is assessed. Running statistics are used in a final step to test whether the reconstruction skill compared to CRU TS 3.10 since 1901 and a set of station temperature observations since 1850 remains robust over time. The question is here whether the relatively small number of analogs is sufficient to reproduce long-term variations and seasonal anomalies realistically over the whole period.

4.2 Data and Methods

4.2.1 Temperature fields of HiResAFF since 1850

The reconstructed temperature fields of HiResAFF since 1961 (Schenk and Zorita, 2012; chapter 2) have been extended back to 1850 (Gustafsson et al., 2012; Meier et al., 2012b) based on the same preselected 22 stations (Fig. 2-2). Table 4-1 gives an overview on the chosen station locations, missing values and updates. With exception of Bodø, the number of missing months is very low. Only in 1850, 4-5 stations are missing. Most stations have been received by Moberg (pers. communication) and consist of updated versions from Jones and Moberg (2003). The time series of the Central England Temperature (CET) is taken from the Hadley Centre (Parker et al., 1992; Parker and Horton 2005) and Vienna from the HISTALP database provided by ZAMG (Auer et al., 2007). Updates have been taken from different sources as indicated in Tab. 4-1. Information about data quality and homogenisation or potential influence of urban heat effects on different stations is given in these references. The well known problem of early industrial warm bias (clear warm bias in summer, partly slight cold bias in winter) in historical temperature observations (Böhm et al., 2010, Frank et al., 2010) and the warm bias in the long Uppsala-Stockholm record (Moberg et al., 2003) is mostly virulent before 1850 and hence is not expected to affect the current reconstruction.

Table 4-1: Station information providing monthly mean station temperature (T2m) used for the reconstruction of HiResAFF 1850-2009. If a second source is given for the data it refers to the update. Source [1] = CRU stations (Jones and Moberg, 2003), [2] = ECA&D (Klein Tank et al., XXX), HadCET = Hadley Centre (Parker et al., 1992), FMI = Finish Meteorological Institute, DWD = German Weather Service, ZAMG = Zentralanstalt für Meteorologie und Geodynamik, Vienna/Austria (Auer et al., 2007). Missing values for each station are given in % for the period 1850 - 2007.

Location (name)	WMO Identifier	Lat. [° N]	Lon. [° E]	Height [m]	Start [year]	End [year]	Added [year]	Misval. [%]	Sources
Arhangelsk	22550	64.32	40.28	13	1850	2001	2009	0.00	1, 2
Bergen (Florida)	1317	60.23	5.20	36	1850	1999	2009	0.00	1, 2
Berlin (Schönefeld)	10385	55.23	13.31	48	1850	2001	2009	0.11	1, 2
Bodo	1152	67.16	14.28	13	1868	1991	2009	11.39	1, 2
CET (~Midlands)*		57.33	1.50		1850	2009		0.00	HadCET
Copenhagen (Ros.)	6117	55.35	12.08	44	1850	2000	2007	0.00	1, DWD
De Bilt	6260	52.06	5.11	4	1850	2001	2009	0.00	1, 2
Haparanda	2196	65.50	24.09	6	1850	2002	2007	0.00	1, DWD
Helsinki (Vantaa)	2974	60.19	24.58	56	1850	2004	2009	0.26	1, FMI
Hohenpeissenberg	10961	47.48	11.01	986	1850	1997	2009	0.00	1, 2
Kiev (Zhulyany)	33345	50.24	30.27	168	1850	2001	2009	2.64	1, 2
Oslo (Fornebu)	1488	59.54	10.37	17	1850	2001	2009	0.00	1, 2
Paris (Le Bourget)	7150	48.58	2.27	65	1850	1995	2009	0.11	1, 2
Prague (Ruzyne)	11518	50.06	14.17	365	1850	2001	2007	7.38	1, NOAA
Riga	26422	56.58	24.04	3	1850	1989	2009	0.90	1, 2
Saint Petersburg	26063	59.58	30.18	4	1851	1991	2009	1.79	1, 2
Stockholm	2464	59.21	17.57	11	1850	2009		0.00	2
Trondheim	1271	63.28	10.56	17	1850	2000	2009	0.63	1, 2
Vardo	1098	70.22	31.06	15	1851	1991	2009	0.69	1, 2
Vienna (H. Warte)	11035	48.15	16.22	200	1850	2008		0.00	ZAMG
Vilnius	26730	54.38	25.17	156	1850	1990	2009	1.90	1, 2
Warsaw (Okecie)	12375	52.10	20.58	107	1850	2001	2007	0.05	1, NOAA

* Coordinates calculated as the centre point of the stations Lancashire, London and Bristol (Parker et al., 1992).

4.2.2 Datasets for Validation

For the evaluation of seasonal temperature variations, trends and their spatial pattern, daily mean temperature fields from the third version of high resolution gridded near-surface temperature dataset of CRU-TS3.10 are used covering the period 1901-2009 (Harris et al., 2013; access via http://www.cru.uea.ac.uk/cru/data/hrg/) with a horizontal resolution of 0.5° x 0.5° (~50 km). To produce a gridded product from different local records, CRU uses a climate anomaly method (CAM, Jones, 1994) where all stations are first reduced to anomalies for the common period 1961-90. After taking an un-weighted average of all station anomalies within a grid box, the climate normals are added in a final step to get the absolute values (Mitchell and Jones, 2005). Based on the station counts used over Northern Europe, a strong increase in the number of stations over time increases the grid point and spatial temperature information towards recent decades.

In a similar way, the much coarser resolved gridded fields of the third version CRUTEM3 (Brohan et al., 2006; access over http://www.cru.uea.ac.uk/cru/data/temperature/ with more information on http://www.metoffice.gov.uk/hadobs/crutem3/) provide un-weighted averages of all station anomalies within a grid box for the period 1850-2010 with a horizontal resolution of 5°x5°. Due to the very coarse resolution, reconstructed trends are directly compared with the trend analysis of BACC (2008) estimating the trends from spatial average seasonal temperatures north and south of 60° N based on CRUTEM3.

4.2.3 Analysis

For the comparison with gridded observations, the higher resolved fields from HiResAFF and RCAX are bi-linearly interpolated on a common grid of CRU TS3.10 with 0.5°x0.5° horizontal resolution. For the period 1961-2007, Pearson correlation and variance ratios ($\phi = \sigma_{\mathrm{Re}c} / \sigma_{Obs}$) of seasonal mean temperatures are calculated as in chapter 2 and field averages of the results are presented in Tab. 4-2. In addition, pattern correlations are calculated for the mean (Pcor(μ)) and variance (Pcor(σ)) (Tab. 4-2) and decadal trends (Pcor(τ)) (Tab. 4-4) between the three datasets for seasonal mean temperatures. Running 31yr-statics of these skill measures are additionally used to estimate whether the reconstruction skill remains stationary over the whole period. To test whether analogs are on average equally easily found within the period 1850-2007, the Euclidian distance between the monthly temperature analogs (Eq. 1) is also presented as time series of RMSE between analogs derived from 22 stations on monthly basis for different seasons (Fig. 4-5).

Trends are estimated by fitting a regression line to the time series using ordinary least-squares (OLS). This is justified as seasonal mean temperature series are not characterized by strong outliers. The latter case would require other methods like "Sen-slope" (Sen, 1968) where the slope between each possible pair of the time series is estimated and the median over all slopes (length n yields ½ n · (n-1) slopes) provides the trend estimate being robust against outliers. Whether trends are statistically significantly different from zero is tested by parametric 2-sided t-test with $p<0.05$. The time series fulfil linearity between the parent distributions and do not exhibit outliers or notable serial correlation ($N_{effective} \sim N$) of year-to-year variations of seasonal mean temperatures. Also residuals show negligible serial correlation. Jones and Moberg, 2003; BACC, 2008) have already shown that the often used non-parametric rank-based Mann-Kendall test (Kendall, 1975) leads to similar results so that it is not tested here.

4.3 Results

4.3.1 Cross-comparison for HiResAFF, CRU and RCAX (1961-2007)

The validation of HiResAFF with the reference model fields RCAX in chapter 2.3.3 proofed relatively high correlations of 0.78 and 0.65 for monthly means for January and July (Tab. 2-1) with an underestimation of variance of around 13% and 17% for these months (Tab. 2-2), respectively. Different to chapter 2, correlations and variance ratios are now calculated on seasonal basis including CRU TS 3.10 as observational reference (Tab 4-2). In addition to the field average of temporal correlations and variance ratios, the spatial pattern correlations for the 1961-2007 means and standard deviations are shown in Tab. 4.2 as cross-comparison between HiResAFF, the ERA40-driven RCAX simulation and CRU TS 3.10.

The results in Tab. 4-2 confirm the high correlations on monthly scale (Tab. 2-1) also for seasonal mean temperatures between all datasets with highest values for winter and lowest agreement in spring. The temporal correlations of RCAX with CRU TS are generally higher than those of HiResAFF with CRU TS. However, it should be also noted that the temporal correlations of HiResAFF are slightly higher with CRU TS than with the reference fields of RCAX. Seasonal variance of CRU TS is underestimated by both, HiResAFF and RCAX, with the latter being closer to CRU than the reconstruction. The loss in variance by the reconstruction compared to CRU TS is smallest with 9% in winter and largest with around 25% in spring and summer.

Table 4-2: Cross-comparison of HiResAFF (REC) with RCAX and CRU TS 3.10 for seasonal means, ratio of seasonal standard deviations and pattern correlation of the mean and standard deviation for the period 1961-2007. In brackets: skill including sea-areas.

Products	Season	Correlation	Pcor(μ)	Rvar	Pcor(σ)
REC-CRU	DJF	0.87	0.62	0.91	-0.61
RCAX-CRU		0.93	0.63	0.97	-0.66
REC-RCAX		0.84 (0.81)	1.00 (1.00)	0.93 (0.93)	0.98 (0.99)
REC-CRU	MAM	0.69	0.04	0.76	-0.64
RCAX-CRU		0.81	0.03	0.80	-0.34
REC-RCAX		0.66 (0.64)	1.00 (1.00)	0.96 (0.88)	0.79 (0.83)
REC-CRU	JJA	0.74	-0.45	0.74	-0.45
RCAX-CRU		0.86	-0.44	0.79	-0.20
REC-RCAX		0.71 (0.66)	1.00 (1.00)	0.95 (0.91)	0.51 (0.69)
REC-CRU	SON	0.78	0.39	0.84	-0.71
RCAX-CRU		0.86	0.39	0.91	-0.59
REC-RCAX		0.70 (0.70)	1.00 (1.00)	0.93 (0.91)	0.93 (0.95)

Tab. 4-3: Sources for the lost variance of HiResAFF (REC) compared to CRU TS 3.10.

Season	REC-CRU absol. deviation	REC-CRU recon. deviation	RCAX-CRU model deviation
DJF	9%	33% (3/9)	78% (7/9)
MAM	24%	20% (4/20)	83% (20/24)
JJA	26%	19% (5/26)	81% (21/26)
SON	16%	44% (9/16)	56% (9/16)

A closer look on the source for the loss in variance of the reconstructed fields is given in Tab. 4-3. Taking the loss of variance of HiResAFF compared to CRU TS as absolute deviation (100%), the relative contributions of the reconstruction method and the RCAX simulation can be estimated. Tab. 4-3 shows that the loss in variance of HiResAFF compared to CRU is mainly caused by taking the fields from the RCAX model. The reconstruction method contributes around 20% to the loss in variance in spring and summer, about one third in winter and 44% in autumn. While in absolute terms spring and summer show the largest deviation in variance relative to CRU, the lowest performance of the method is in autumn.

In addition to temporal correlation and variance, Tab. 4-3 also shows spatial pattern correlations for seasonal means and standard deviations. The spatial pattern of seasonal mean temperatures between CRU TS and HiResAFF or RCAX agree well in winter and reasonably well in autumn while no agreement is found for spring and negative pattern correlation in summer. The perfect reproduction of the pattern of means by HiResAFF relative to RCAX shows that the reconstruction skill is here strongly dependent on the used model fields which are only re-sampled in time by the analog-method.

The spatial dependency of HiResAFF on the used RCAX fields is to some extend also found for the variance pattern. Here, the spatial pattern correlations of seasonal standard deviation of HiResAFF and RCAX show generally negative values compared to CRU TS. The reconstruction itself yields very high positive spatial correlations relative to the used RCAX fields in winter and autumn (>0.9) with lower agreement of 0.51 in summer. The positive pattern correlation for standard deviation between HiResAFF and RCAX shows that the deviation from CRU is caused by the RCAX fields rather than the reconstruction method itself. However, different to the perfect reproduction of the spatial pattern of mean seasonal temperatures, the analog-method leads to deviations in the pattern of variability relative to RCAX i.e. in summer.

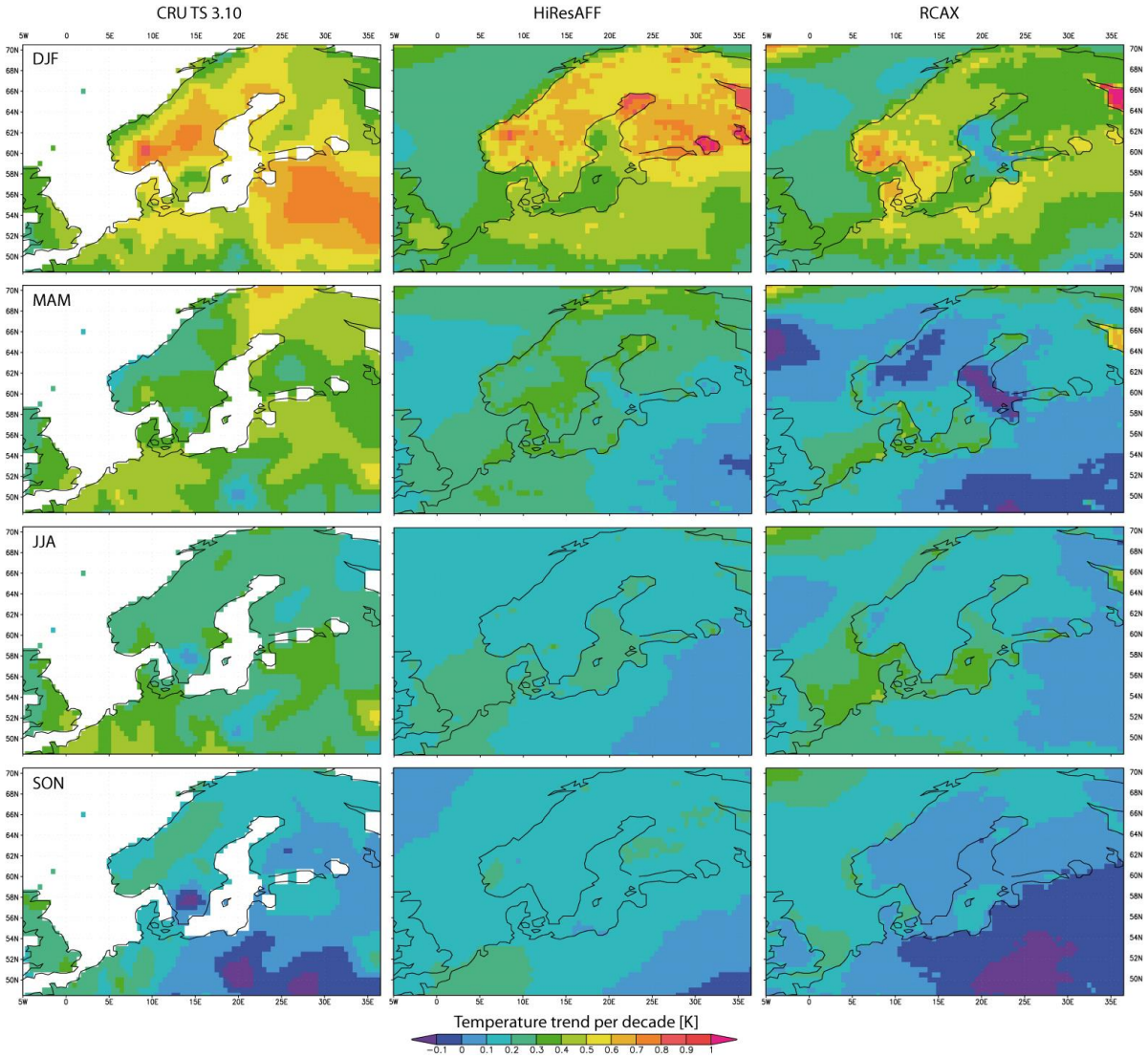

Fig. 4-1: Comparison of decadal near-surface temperature trends [K per decade] calculated over the period 1961-2007 for different seasons.

Another aspect besides temporal correlation and variability relates to a realistic reproduction of potential long-term changes such as seasonal warming trends. The reconstructed seasonal temperature trends are shown in Fig. 4-1 in comparison with CRU TS and RCAX for the period 1961-2007. Figure 4-2 shows the regional differences in the magnitude of trends between the different datasets. As a quantitative measure for the comparison of trends, Tab. 4-4 presents the field average of the squared difference of decadal trends (diff. sqr.) and the pattern correlation of trends between the different datasets.

The regional temperature trends in Fig. 4-1 show generally a better agreement for the strength of the trends between reconstruction (REC) and CRU TS than between RCAX and CRU TS. This is also reflected in a smaller error for the reconstruction regarding the averaged squared difference of the trends in Tab. 4-4. In contrast, the pattern correlation of the trends shows a better agreement between CRU and fields from the regional climate model RCAX than between CRU and REC.

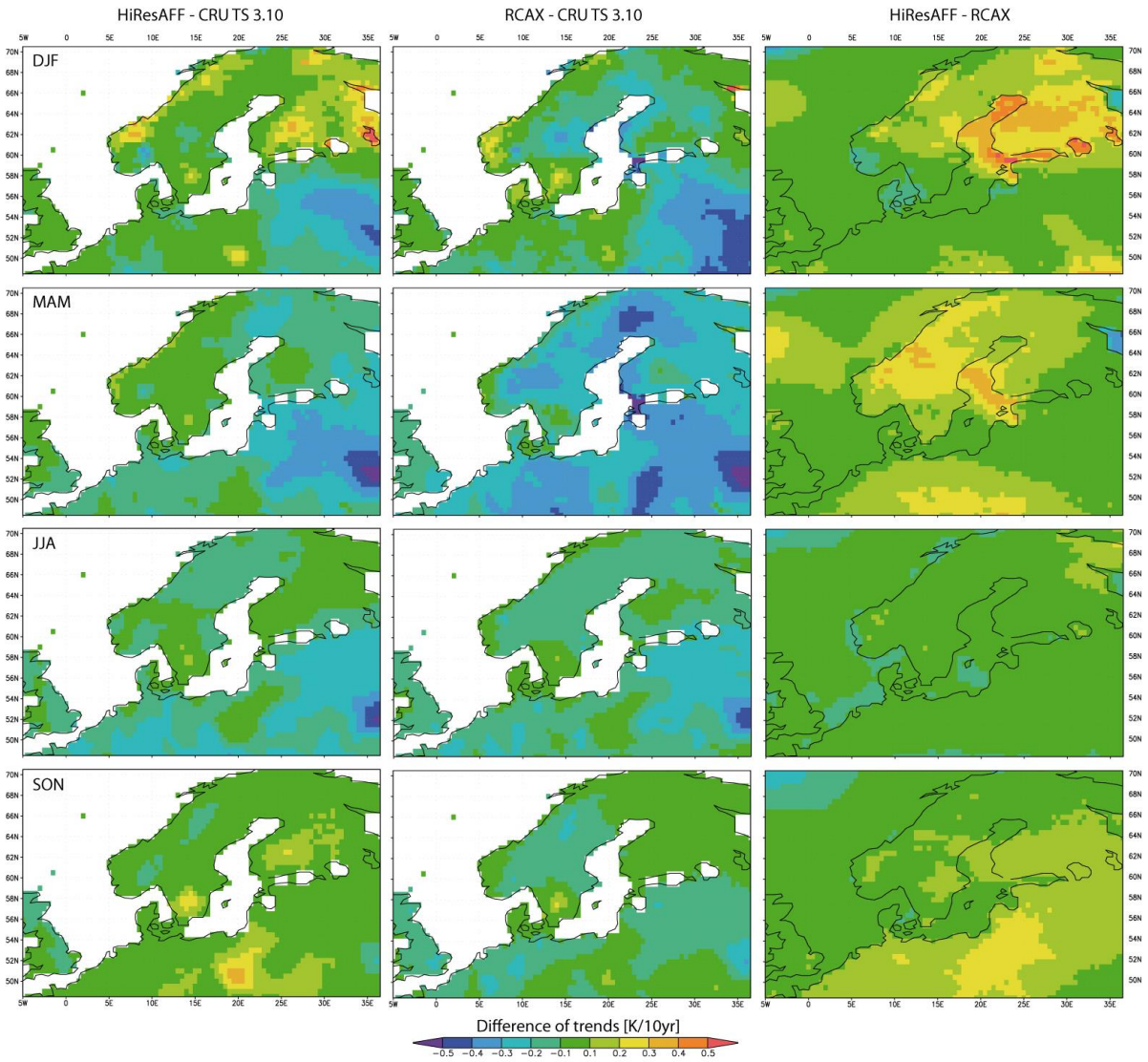

Fig. 4-2: Differences of decadal trends between the different datasets. Blue colours indicate an underestimation of trends compared to the reference of CRU TS (first two columns) or the RCAX simulation (last column) while yellow to red indicates an overestimation.

Tab. 4-4: Squared differences (diff sqr) of decadal trends and the pattern correlation of trends (pcor) between different datasets for land T2m based on annual and seasonal means 1961-2007. The last column includes also trends over sea.

Season	**REC-CRU**		**RCAX-CRU**		**REC-RCAX**		**REC-RCAX (+ sea)**	
	diff sqr	Pcor(τ)	diff sqr	Pcor(τ)	diff sqr	Pcor(τ)	diff sqr	Pcor(τ)
ANN	0.0124	-0.321	0.0333	0.208	0.0086	0.548	0.0104	0.738
DJF	0.0258	-0.509	0.0389	-0.390	0.0209	0.657	0.0236	0.568
MAM	0.0367	0.061	0.0843	0.261	0.0187	0.628	0.0210	0.648
JJA	0.0263	0.311	0.0261	0.454	0.0039	0.690	0.0028	0.630
SON	0.0074	0.002	0.0130	0.666	0.0131	0.677	0.0173	0.414

As depicted in Fig. 4-1, the temperature trends over 48 years in CRU TS shows the strongest warming over the SE domain and central Scandinavia in winter and northernmost Scandinavia in spring. Moderate warming trends exist in summer and little warming and even a slight cooling trend in the SE domain took place in autumn. The RCAX simulation shows generally a weaker warming without any pronounced warming over the SE in winter and spring. The difference in the magnitude of trends is strongest in spring (Fig. 4-2) with a clear underestimation of warming trends for all regions.

The warming trend of the reconstruction is more realistic in winter and spring albeit with spatial differences (leading to negative pattern correlation in Tab. 4-4) showing also no warming over the SE. Instead, HiResAFF overestimates warming over Fenno-Scandinavia compared to both, CRU TS and RCAX. The agreement with warming trends of CRU TS in spring is also considerably higher with the reconstruction while the simulation shows almost no and partly even cooling trends compared to warming trends in observations. Warming trends in summer are underestimated for HiResAFF and RCAX. Slight negative trends in the SE domain in CRU TS are also shown in the reconstruction and RCAX with too strong cooling in the model and too little cooling in the reconstruction.

4.3.2 Temperature trends of HiResAFF compared to CRU TS 1901-2009

The aim of the previous section has been to estimate the agreement of seasonal temperature trends including their spatial pattern between different datasets independent from the question whether these trends are relevant given the high variability and a relatively short period of only 48 years. Extending the analysis to the full period covered by CRU TS 3.10 since 1901, the length of now 109 years allows a slightly more robust trend estimation of seasonal temperatures.

The seasonal trends per century for the period 1901-2009 are shown for CRU TS and HiResAFF in Fig. 4-3 together with absolute differences of these trends (right column). While the strong warming trend of CRU in winter and spring found for the period 1961-2007 also persists since 1901 over the SE domain, clear long-term warming trends are mainly visible for spring while winter shows little and in northernmost regions no warming since 1901. The difference of reconstructed trends compared to CRU is relatively low (±0.2 K/100a) for spring with exception of the SE-domain while the reconstruction shows generally stronger winter warming trends again with exception of the SE-domain. The weak warming trends in summer are reconstructed with relatively similar values but spatially more heterogeneous. While warming trends of CRU in autumn are relatively similar to summer, the reconstruction shows an overestimation from Denmark over Scandinavia of around +0.3 to +0.6 K/100a.

A comparison of temperatures trends averaged over the whole land area in Tab. 4-5 shows a relative good agreement for annual and winter mean temperatures. The reconstruction clearly underestimates warming trends in spring and summer while they are too strong in autumn. The negative pattern correlation inherited by RCAX for annual and winter trends (Tab. 4-4) is maintained by the reconstruction also for the period 1901-2008/09 while a considerable improvement is achieved for other seasons. Also the average of the squared differences of decadal trends in Tab. 4-5 shows an improvement compared to the shorter period since 1961. Despite regional and intra-seasonal differences in the reconstructed trends, HiResAFF is able to realistically reproduce the magnitude of the overall warming trends for annual mean temperatures since 1901.

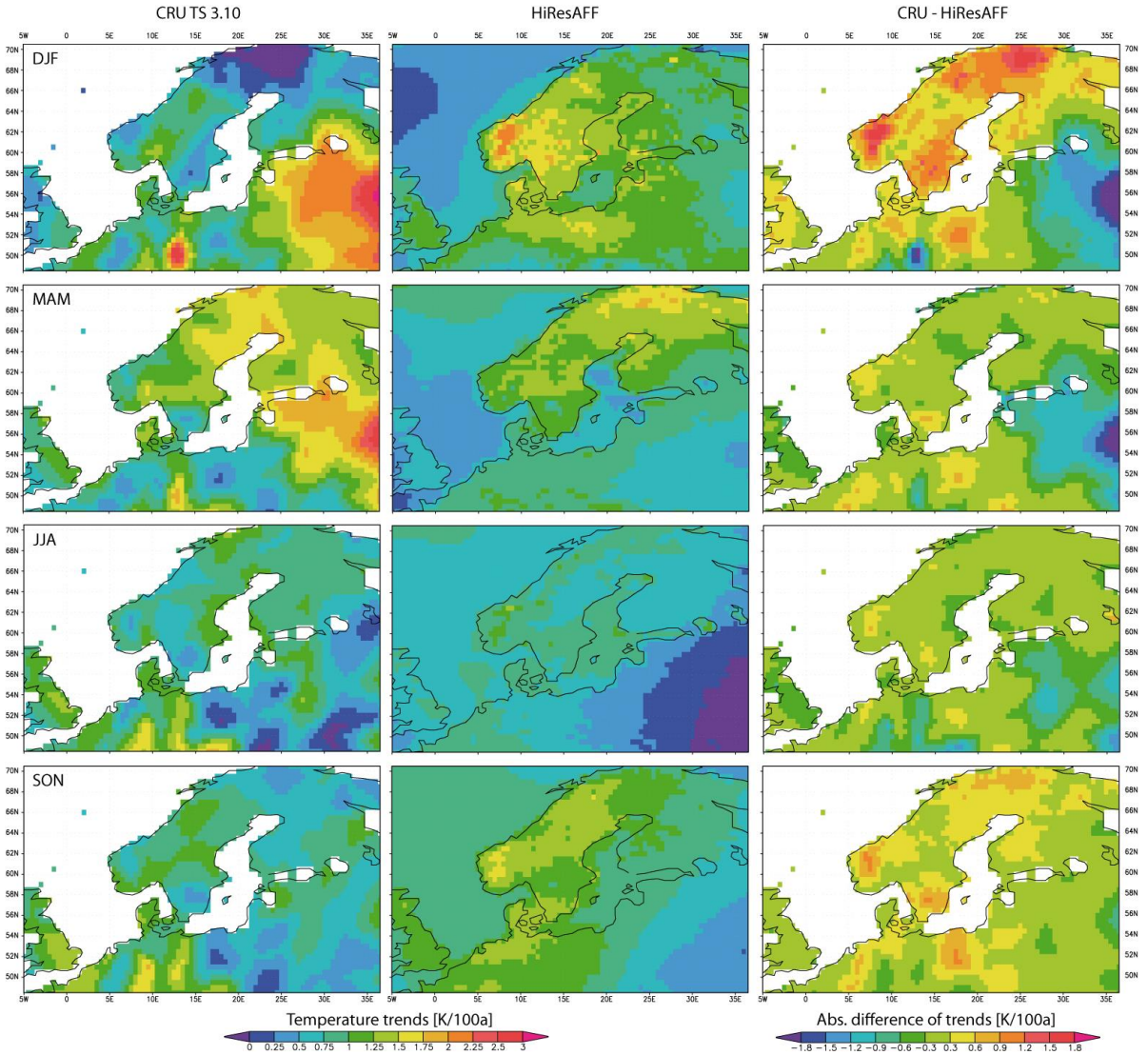

Fig. 4-3: Comparison of centennial near-surface temperature trends [K per century] in the period 1901-2009 and absolute difference of trends HiResAFF minus CRU TS [K per century] (right column) calculated for different seasons.

Tab. 4-5: Centennial trends of seasonal mean temperatures for 1901-2009 and annual means for 1901-2008, their spatial pattern correlation and squared difference of trends.

Season	Trends CRU [K/100a]	Trends REC (+ sea) [K/100a]	Pcor	Avg. $(\Delta\tau)^2$
ANN	+0.939	+0.844 (+0.778)	-0.323	0.0018
DJF	+1.043	+1.099 (+0.908)	-0.200	0.0059
MAM	+1.229	+0.962 (+0.840)	0.177	0.0030
JJA	+0.780	+0.535 (+0.587)	0.500	0.0016
SON	+0.801	+0.946 (+0.917)	0.469	0.0012

With a length of 109 years, one may question whether seasonal temperature trends are significantly different from zero. This is estimated for land-temperatures using a 2-sided t-test with the result of little significant trends in winter over most areas (not shown). Due to the very high variability and dependent on the region, warming trends are statistically not significant ($p<0.05$) below 0.5 to 0.6 K per century over the period 1901-2009 in winter. Based on CRU, significant warming in winter is therefore limited to the SE domain, eastern UK and coastal areas around the British Channel plus parts of SE-Germany to Czech

Republic. HiResAFF shows significant warming for the whole British Islands, North Sea coastal areas and Scandinavia south of around 64° N including the central, southern and western Baltic Sea and Poland. In contrast to winter, warming trends in spring are significant over almost all regions (95%-level around 0.5 K/100a) with exception of areas towards SE. With lower magnitudes mostly not exceeding +1 K/100a, this is also the case for summer (95%-level around 0.3 K/100a). In autumn (95%-level around 0.5 K/100a), a meridional pattern exists with significant warming in the west towards no significant trends in the east with HiResAFF showing stronger trends and hence larger areas i.e. over northern Scandinavia and Finland with still significant warming of up to +1 K/100a.

4.3.3 Temperature trends for the period 1850-2009

The trend analysis in the previous part proved that with regional and intra-seasonal differences, HiResAFF is principally able to reproduce long-term changes of temperatures since the beginning of the 20th century. The reconstructed seasonal trends for the full period 1850-2009 are shown in Fig. 4-4 with white areas representing non-significant trends ($p<0.05$). Table 4-6 shows corresponding centennial trends of seasonal and annual mean temperatures for the weighted average of the whole field and separated for land and sea areas, respectively. In addition, trends for different sub-regions of the Baltic Sea (70° N – 50° N; 10° E – 34° E) north and south of 60° N are given.

A comparison of the terrestrial temperature trends since 1901 (Tab. 4-5) with those since 1850 (Tab. 4-6) shows a notable smaller non-significant warming rate for summer of +0.18 K per century since 1850 compared to +0.54 K estimated since 1901 over Northern Europe. With exception of winter, also spring, autumn and annual long-term trends since 1850 are slightly lower by around 0.1 K per century than since 1901.

As can be expected from the high heat capacity of water, near-surface atmospheric warming trends are generally lower over sea areas with exception of summer (Tab. 4-6). The stronger warming trends over sea in summer are consistent with increased heat storage due to strong warming trends in spring. Warming over the Baltic Sea region since 1850 is now considerably stronger north of 60° N in all seasons while this was only the case in spring since 1901 based and CRU TS3.10 and HiResAFF (Fig. 4-3).

A comparison of trends on regional basis with gridded observations like CRUTEM3 is hampered by the very coarse resolution of 5°x5° and missing grid information for northern and eastern grid points in the early period of the 1880s. A comparison with warming trends calculated by Jones and Moberg (2003) and updated for BACC (2008) for the period 1871-2004 shows a very high agreement for seasonal and annual trends over the Baltic Sea region north and south of 60° N (Tab. 4-7). The only exception is again the underestimation of warming trends in spring similar to findings since 1901 and since 1961 for the RCAX simulation. The statistical non-significance of warming trends (+0.9 K/100a) found over the Baltic Sea region north of 60° N in winter since 1871 indicates that the trends are not very robust due to the high variability i.e. in winter. This is also confirmed by relatively large standard errors (SE, standard deviation of the residuals between original time series y and trend line y_t) of the trends calculated for the period 1850-2009 (not shown).

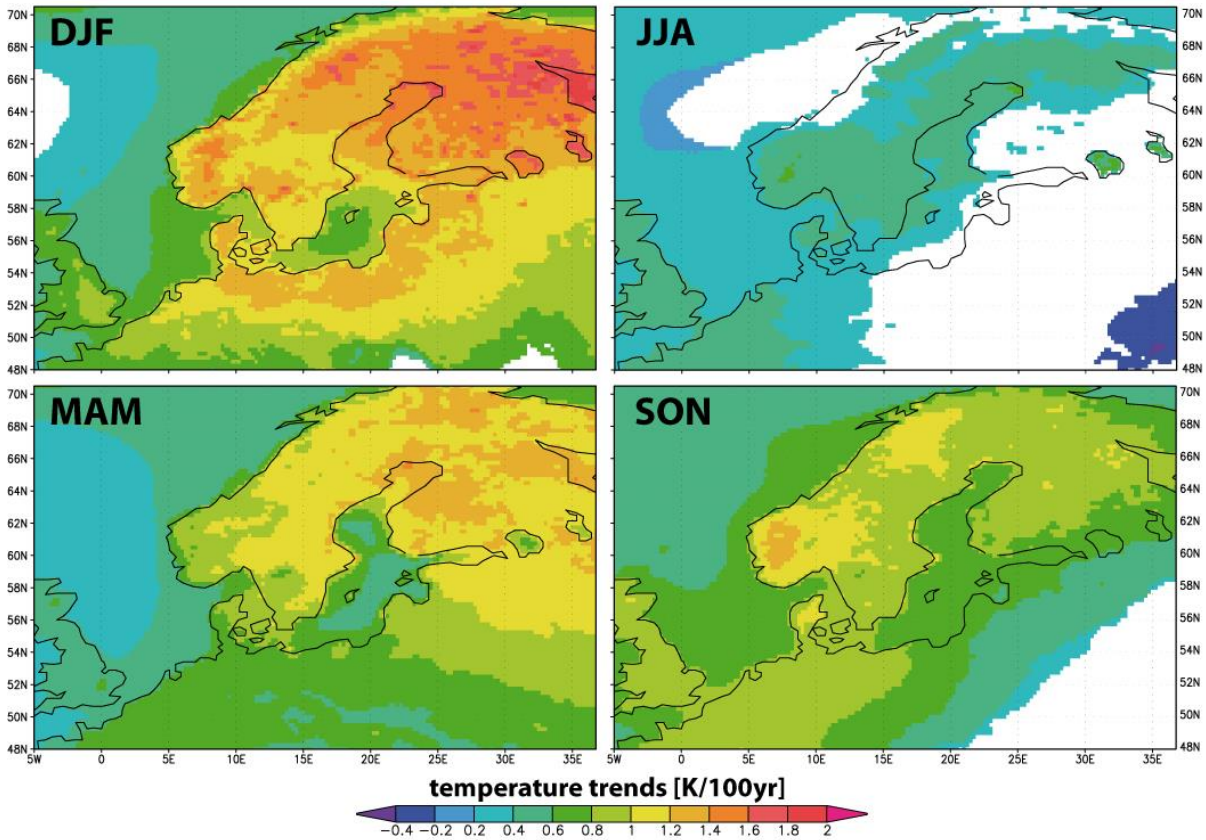

Fig. 4-4: Reconstructed centennial trends of seasonal mean temperatures over the period 1850-2009. Non-significant trends based on a 2-sided t-test with p<0.05 are masked by white areas.

Tab. 4-6: Weighted field average of reconstructed centennial trends of seasonal and annual mean temperatures over different regions, land and sea-areas for 1850-2008/09. Significant trends using a 2-sided t-test with p<0.05 are in bold.

Season	All	Land	Sea	BS region	BS North	BS South	Baltic Sea
ANN	**0.629**	**0.698**	**0.534**	**0.737**	**0.920**	**0.613**	**0.701**
DJF	**0.942**	**1.117**	**0.698**	**1.213**	**1.377**	**1.001**	**1.092**
MAM	**0.756**	**0.895**	**0.564**	**0.962**	**1.154**	**0.833**	**0.803**
JJA	**0.223**	*0.184*	**0.277**	*0.174*	**0.367**	*0.045*	**0.332**
SON	**0.685**	**0.704**	**0.660**	**0.716**	**0.909**	**0.587**	**0.741**

Tab. 4.7 Comparison of centennial trends over the Baltic Sea region north and south of 60°N for the period 1871-2004 based on CRUTEM3 (Jones and Moberg, 2003; BACC, 2008) and HiResAFF. (Non-)significant trends according to 2-sided t-test with p<0.05 are in italics/bold.

Season	CRUTEM3	HiResAFF	CRUTEM3	HiResAFF
	BS North of 60° N		BS South of 60° N	
ANN	**+1.0**	**+0.9**	**+0.7**	**+0.6**
DJF	*+0.9*	*+0.9*	**+1.0**	**+1.0**
MAM	**+1.5**	**+1.2**	**+1.1**	**+0.8**
JJA	**+0.6**	**+0.5**	*+0.3*	*+0.1*
SON	**+0.8**	**+0.9**	**+0.6**	**+0.6**

4.3.4 Seasonal temperature variations since 1850

The previous comparison with CRUTEM3 confirms a realistic reconstruction of long-term trends since 1871. Concerning the time evolution of the under laying temperature series, the seasonal mean temperature anomalies calculated as a simple average of the 21 predictor stations (without Archangelsk 64° 32' N, 40° 28' E) are used as a benchmark for the reconstructed time series of corresponding closest grid points to these stations. The time evolution of the 21 grid points is very similar to those of the field average (not shown) due to the spatially well distributed locations and the generally high co-variability of seasonal temperature variations. All time series are seasonal anomalies relative to the mean of 1961-1990 hereby removing the systematic mean bias of the different datasets.

The correlations for the different seasons over the period 1850-2007 are generally very high (r>0.86) with slope parameters very close to 1 for the regression line in Fig. 4.5. However, the scatter plot also indicates notable deviations for the extreme seasonal anomalies where HiResAFF shows a systematic underestimation of the warmest anomalies (scatter always above the regression line) compared to observations. Large deviations for the coldest anomalies exist i.e. in spring indicated by the diffuse scattering around the regression line.

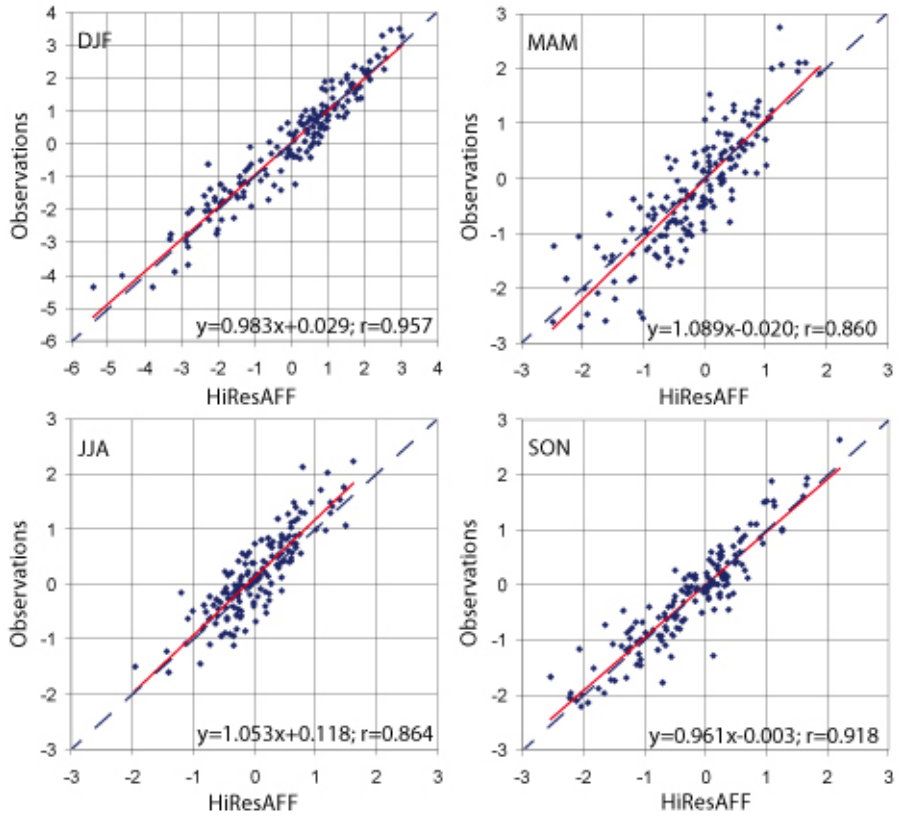

Fig. 4-5: Scatter plot of seasonal mean temperature anomalies derived from station observations and closest grid points from HiResAFF over the period 1850-2007. Additionally shown are the regression lines (red) and the corresponding slope parameters, residuals and correlations.

The time series in Fig. 4-6 confirm the high agreement in phase (correlation) with a slight underestimation of variability and magnitudes. Also the cold bias/underestimation of the warming trends over the last decades as in Fig. 4-1 is visible for HiResAFF and RCAX. Coldest spring temperatures are underestimated leading to a relative warm bias before around 1900, where observations show partly very cold conditions. Both, the underestimation of the last warm decade and the cold decades before 1900 contribute to the underestimation of the warming trends in spring and explain the difference compared to CRUTEM3 (Tab. 4-7).

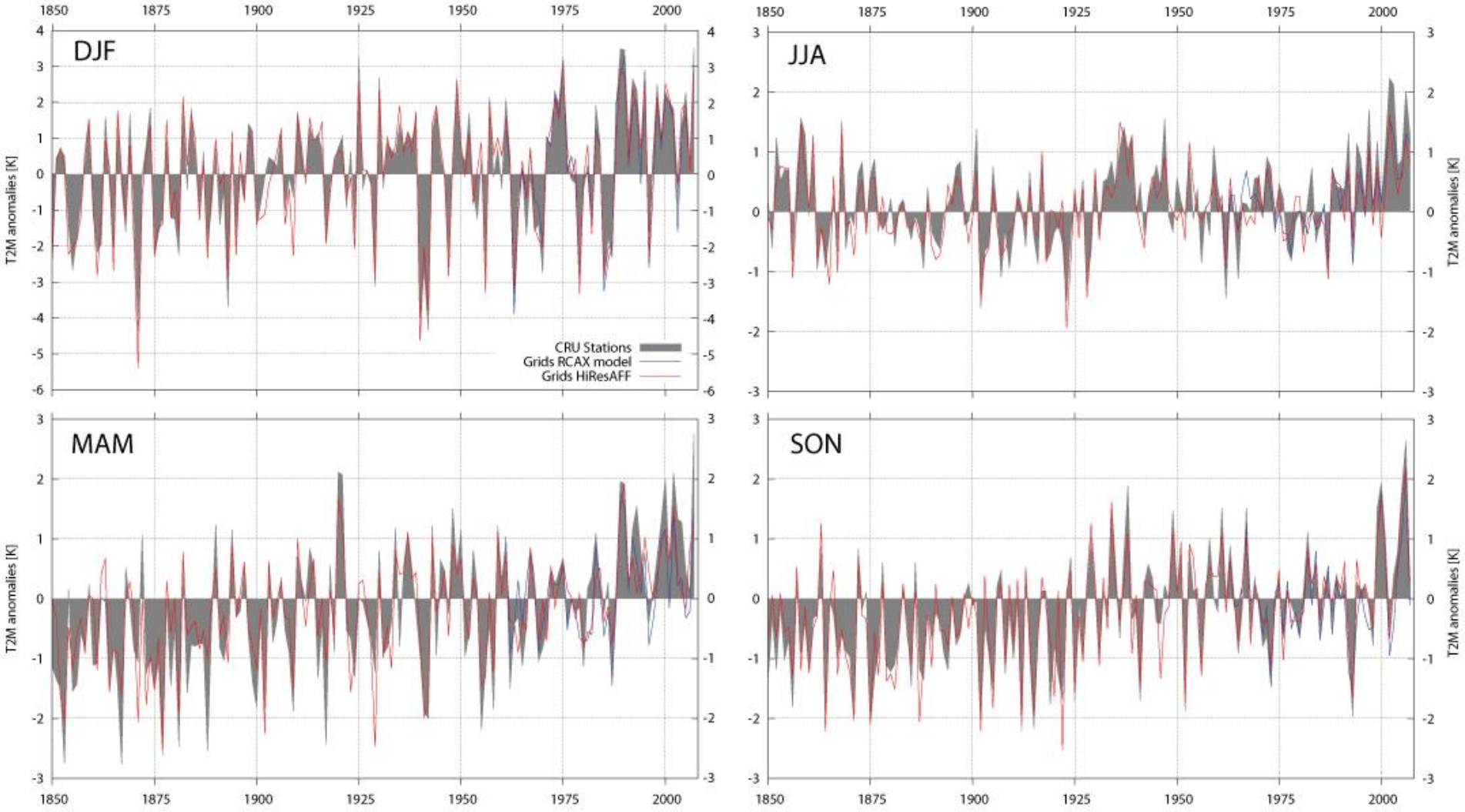

Fig. 4-6: Seasonal mean temperature anomalies 1850-2007 calculated as average over the N=21 used predictor stations (grey) compared to the average from the closest grid points from HiResAFF (red) and the RCAX model (blue) since 1961. Anomalies are relative to mean of 1961-1990.

4.3.5 Robustness of the reconstruction skill over time

The temporal agreement of time series in Fig. 4-6 do not suggest a considerable change in the reconstruction skill over time compared to observations. To quantify the temporal robustness, the retrieved Euclidian distance between the found monthly analogs, which is based on the 22 predictor stations, provides a first estimation whether analogs are on average equally easy found over time. Figure 4-7 presents seasonal time series of the distance between the found monthly temperature analogs as anomalies of the RMSE relative to the corresponding mean RMSE of 1850-2007.

The trend lines for the RMSE show a slight decrease in pattern similarity estimated from the used 22 stations back in time. This is most pronounced for winter months before around 1900 as indicated by the running 30-monthly mean remaining above the long-term mean RMSE. A right-centred running mean is used here as the early years in 1850 contain five, 1851-1953 three and 1854-1867 two missing values. The slight decrease of the RMSE over time in spring and summer can be explained by smaller RMSE values in recent decades since 1970 (MAM) and 1945-2000 (JJA) compared to the long-term mean RMSE. With exception of winter before 1900, the RMSE does not show a strong change in the distance of analogs consistent with the high co-variability in Fig. 4-6.

As the distance between analogs is not directly related to the reconstruction skill, Fig. 4-8 shows additionally the temporal changes in 31-yr running correlations, variance ratio and mean bias which together contribute to the RMSE of HiResAFF. Again, time series of the average of the 21 stations and closest grid points from the reconstruction are used as anomalies to the long-term mean 1850-2007. The running correlations show on average an increase for DJF, a slight decrease for MAM and JJA and no change for SON from 1850 to 2007. The drop of correlations in winter before around 1920 corresponds to the higher RMSE of the analogs in Fig. 4-7 before around 1900.

The variance ratio indicates on average a more similar variability of the reconstruction relative to observations back in time i.e. compared to the last around 50 years. Comparing the reconstructed 31yr running standard deviations e.g. of DJF with those of observations yields a high co-variability in time (not shown). The decrease of the variance ratio for DJF can be explained by a stronger positive trend for the variability in observations than in the reconstruction (not shown). It is however difficult to estimate the potential effect of data quality in observations on the comparison with the reconstruction. Better and i.e. more frequency temperature measurements could e.g. increase the variability in recent decades in observations while such an effect is not present in the reconstruction. A detailed analysis for the systematic divergence in the variability is however beyond the scope of this work.

With exception of spring, the mean bias does not exhibit a trend over the whole period although different periods show cold and warm bias. All seasons show an increasing cold bias compared to observations for recent decades due to the underestimation of the warming trends noted earlier also for the RCAX simulation. Spring additionally shows an increasing warm bias before around 1900 which together with the cold bias in recent decades explain the strong underestimation of long-term warming trends in spring.

Reflecting the different contributions of temporal changes in correlation, variance and bias, the RMSE decreases on average for all seasons with exception of an increase in summer. This indicates a slight improvement of the reconstruction over time compared to the set of 21 stations. It is unclear, whether quality changes in observations might interfere this result.

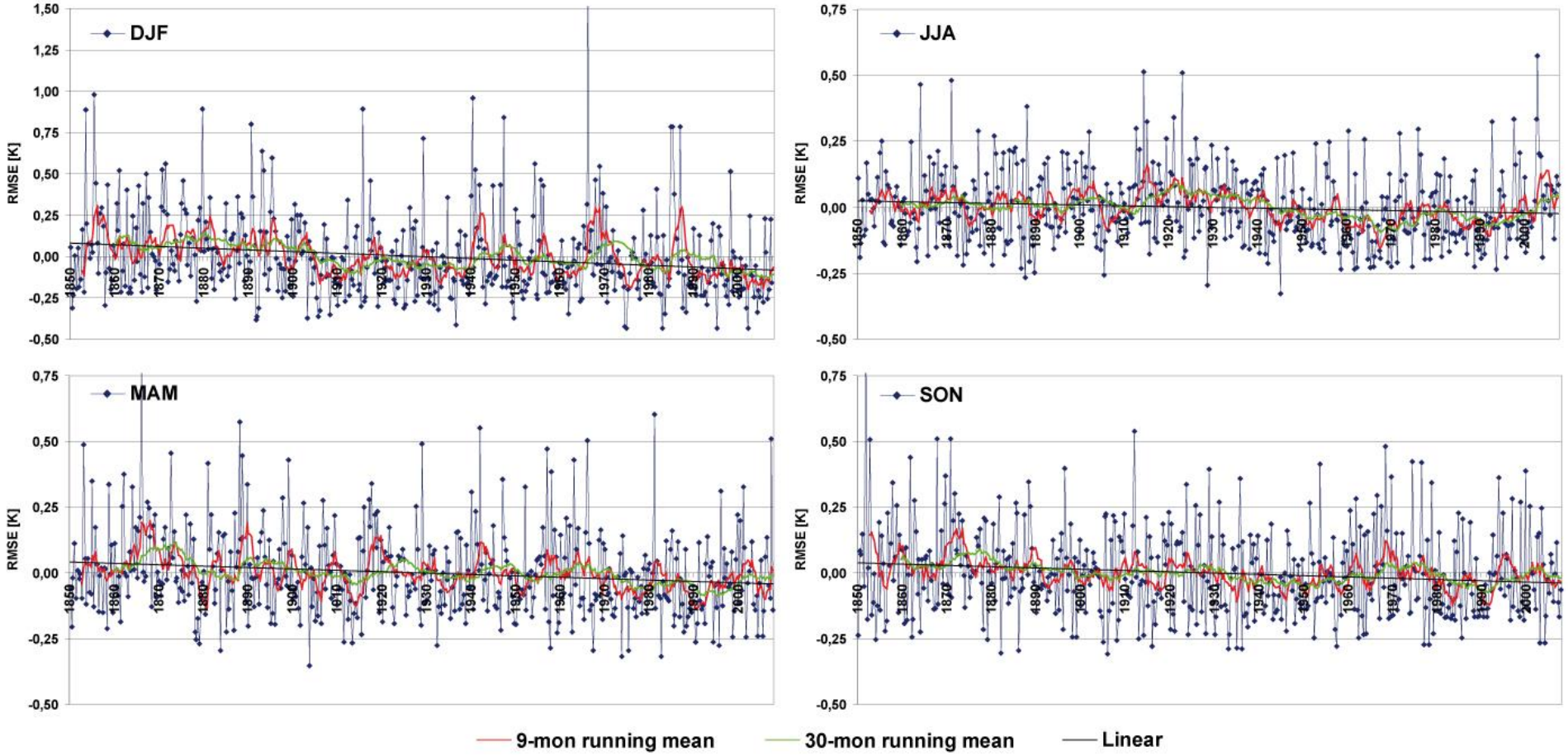

Fig. 4-7: Deviation of the RMSE from the 1850-2007 mean for monthly analogs of each season as derived from the analog-search (chapter 2). Additionally shown are the linear trend and right-centred 9-mon and 30-mon running means to highlight potential long-term changes.

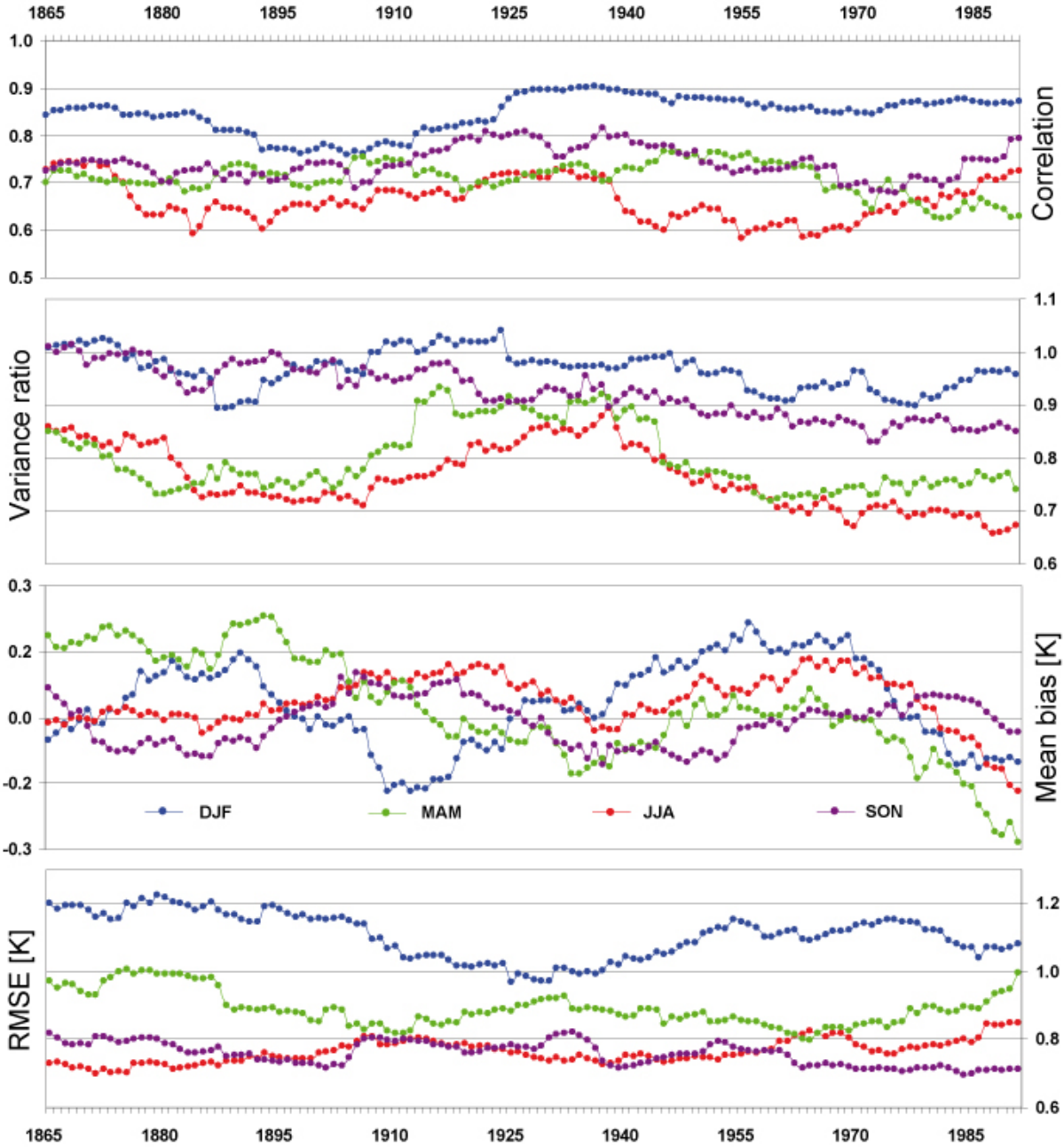

Fig. 4-8: Running 31yr-statistics showing temporal changes of the reconstruction skill between the 21 stations used as predictor and closest grid points from HiResAFF for seasonal mean temperatures over the period 1850-2007. The comparison is based on anomalies relative to the 1850-2007 seasonal means.

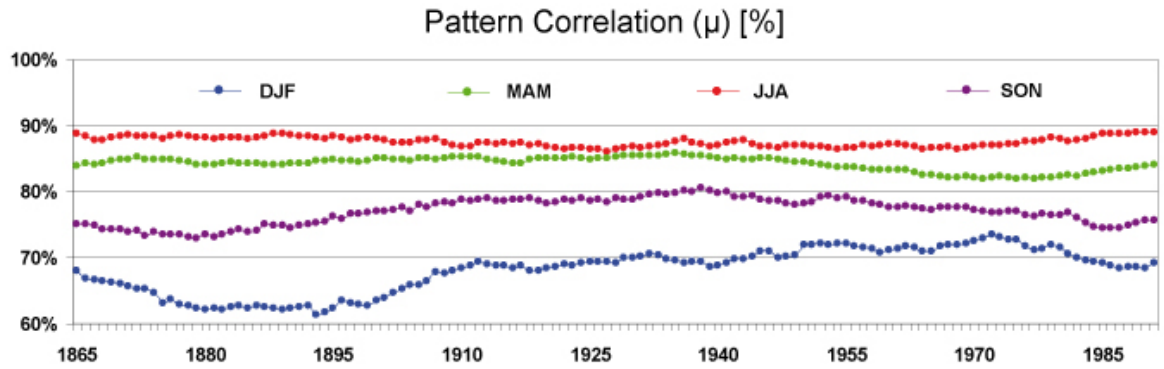

Fig. 4-9: Pattern correlation [r^2, %] between 31yr-means of 21 predictor stations and closest grid points from HiResAFF for different seasons.

4.4 Discussion

4.4.1 Reconstruction skill and model dependency

An important aspect of the analog-upscaling approach generally relates to the question to which extent the reconstruction skill in comparison to observations are dependent on specific properties and potential limitations of the model fields used as target fields. The cross-comparison of HiResAFF, RCAX and CRU TS 3.10 shows a generally good agreement concerning seasonal correlation (Tab. 4-2). An important result is here that the seasonal co-variability of the reconstruction is slightly higher with CRU observations than with the model RCAX. This can be explained by the predictive skill contained in the 22 stations. Over the analog-search, these predictors bring the temporal variations of reconstructed fields slightly closer to the observational dataset than to the "mother fields" of RCAX. This result may very likely be model dependent owing to the gap when linking real observational data to model fields (Schenk and Zorita, 2012; Fig. 2-3, chapter 2.3.2). An improvement for a new reconstruction regarding the gap between observations and model simulation may be achieved by using fields of a regional climate model where spectral nudging was used for the downscaling of reanalysis data (chapter 2.4.1).

Another typical property of model simulations relate to an underestimation of variance compared to observations. This is also the case for RCAX with an underestimation of up to 20% in spring and summer (Tab. 4-2) compared to CRU. The analog-method slightly increases the underestimation of the reconstructed variance although the method only contributes about one fifth to one third of the absolute deviation of the reconstructed mean seasonal variance relative to CRU with exception of a stronger underestimation in autumn. This result confirms in principal a good reproduction of variability by the analog-method while the main deviation from observations is here again dependent on the used model.

Spatial characteristics of the model fields impose a complete dependency on the reconstructed temperature pattern of the seasonal long-term means. This is an expected result of re-sampling fields in time by the AM without affecting the long-term spatial pattern. This effect is partly lower for the spatial pattern of mean seasonal standard deviations where the reconstruction shares only a spatial correlation of 0.51 with the reference fields of RCAX in the example of summer (Tab. 4-2). Nevertheless, the pattern correlation of the reconstruction compared to CRU is to a large extent set by the used RCAX fields. The very low or even negative spatial correlations in two out of four seasons for seasonal means and generally negative pattern correlations for variance are all inherited by RCAX and cause a limited usability in terms of spatial pattern analysis also for the reconstruction. This is however model dependent and also uncertainties in the realism of the spatial patterns of gridded observations like CRU may lead to the problematic pattern correlations.

The ability of HiResAFF and RCAX to reproduce seasonal trends for the period 1961-2007 in agreement with CRU observations (Fig. 4-1) differs considerably. The reconstruction shows a clearly better agreement for the magnitude of trends (Fig. 4-2) leading to smaller values of the averaged squared differences of trends compared to CRU (Tab. 4-4). In contrast, RCAX achieves a better spatial pattern correlation for trends. The RCAX simulation tends to underestimate warming trends in all seasons with a very strong underestimation in spring. It is important to note that the analog-method is able to partly overcome the deviations in the magnitude of trends by RCAX through the temporal re-sampling according to station observations and hence a slightly optimized selection of fields.

Despite the generally high spatial correlation of reconstructed trends with RCAX, Fig. 4-2 shows also that the analog-method is able to deviate from the trend pattern of RCAX i.e. in winter and spring. The reconstruction's ability to deviate from the RCAX trend patterns and even improve the realism of the trend magnitudes is highly important to confidently reproduce spatial and temporal long-term changes regarding trends. The agreement with CRU TS regarding the reconstructed trends increases for the longer period 1901-2009 compared to 1961-2007. Two reasons may explain the improvement: The first is, that trends are now more robust over 109 years while small deviations over 48 years will lead to a larger deviation of retrieved trend estimates. The second explanation could be related to the longer sampling period. With more degrees of freedom, the predictive skill of the 22 stations can now better re-sample the target fields according to observations hereby reducing the deviation of the RCAX fields found for the 48 years.

The scatter plot (Fig. 4-5) and time series of seasonal temperature anomalies of HiResAFF and station temperatures from the period 1850-2007 (Fig. 4-6) show that the strong underestimation of warming trends in spring is a result of RCAX not providing cold or warm enough samples for the strongest seasonal anomalies. This leads to an underestimation of warming trends i.e. in the last decades and also an underestimation of the very cold periods before around 1900 (Fig. 4-6). Variations of running 31yr mean seasonal bias (Fig. 4-8) in spring clearly reflect this effect as a cold bias in recent decades and a relative warm bias prior to around 1900.

The reconstructed underestimation of warming in spring is here a result of two effects: First, the RCAX already strongly underestimated the warming trend in spring so that the predictors of monthly mean station observations are linked to too cold fields of RCAX in recent decades. This effect is principally unknown by the analog-method as the pattern similarity based on the 22 predictors is independent from the realism of the related target fields of RCAX. However, the in comparison to RCAX better agreement of reconstructed trend magnitudes in other seasons show, that the reconstruction is at least partly able to correct the underestimation of trends by RCAX. Therefore, a second accompanying effect seems to play a role here which relates to the relatively small sample of available analogs providing a too narrow distribution for strong seasonal anomalies by RCAX. Both, a wider range and a higher sample size of analogs could at least partly help to overcome the underestimation of warming trends by RCAX as shown for more realistic magnitudes of reconstructed trends in other seasons.

The reconstructed trends over the period 1901-2009 show on average clearly smaller errors compared to CRU TS than for the period 1961-2007. The average of squared differences in Tab. 4.5 is at least an order of magnitude smaller. However, the magnitude of warming trends in spring and also summer are clearly underestimated compared to CRU TS since 1901 while the agreement is very high for seasonal trends of 1871-2004 compared to CRUTEM3 over the Baltic Sea catchment (chapter 4.3.3, Tab. 4-7) with an underestimation of spring warming.

The evaluation of the temporal robustness of the reconstruction skill suggests an overall slight improvement of the reconstruction in terms of the 31yr running RMSE (Fig. 4-8) over the period 1850-2007. Whether the higher divergence in variance of the reconstruction from observations since around 1950 also reflects improvements in observational data quality (Fig. 4-9b) cannot be assessed here. The relatively small long-term change in the reconstruction skill, the only small increase of the distance between analogs (Fig. 4-7) and finally a very close reconstruction of warming trends for 1871-2004 suggest that HiResAFF provides a relatively robust temperature reconstruction for the period since 1850.

4.4.2 Seasonal long-term trends over Northern Europe

Long-term variations and trends of annual or seasonal temperatures have been intensively studied on regional scale based on many long historical observations for Northern Europe and the Baltic Sea region (BACC, 2008). An interesting result of the seasonal trend analysis of CRU TS and HiResAFF for the period 1901-2009 is that with limited regional exceptions, no statistically significant ($p<0.05$) warming took place over Northern Europe in winter (not shown). On the one hand, this is a result of relatively high significance levels due to the large inter-annual and long-term variability in winter and on the other hand the fact, that the period until the 1930-40s has been relatively warm in terms of seasonal anomalies relative to the mean of 1960-91 (Fig. 4-6). Independent from the question of statistical significance, observations even suggest no warming at all over the most northern part of Scandinavia since 1901. Even for the period 1871-2004, the null hypothesis of no significant trends cannot be excluded at a confidence level of 95% for a warming rate of +0.9 K per century over the Baltic Sea region north of 60°N consistent with BACC (2008) (Tab. 4-7).

Different to the period 1961-2007, the strongest warming trends since 1901 or earlier are observed and reconstructed for spring instead of winter. In agreement with the phenomenon of "polar amplification" (cf. Serreze and Barry, 2011 for a review), the region north of 60° N shows clearly stronger warming trends for all seasons than more southward. The region south-east of the Baltic Sea does not exhibit significant warming trends since 1850 in contrast to the north. Based on GISS temperatures for the period 1960-2009, Serreze and Barry (2011) find the strongest polar amplification in winter consistent with CRU TS and HiResAFF in Fig. 4-1. However, the weak (HiResAFF) or even absent winter warming (CRU TS) in most northern regions since 1901 (Fig. 4-3) also shows that this effect is not continuous and undergoes spatiotemporal and intra-seasonal shifts e.g. towards spring. An analysis of temperature records north of 62° N by Polyakov et al. (2002) even suggests that over the long period 1875-2001, no polar amplification can be stated relative to the warming of the northern hemisphere. This appears to be different for the Baltic Sea region (Fig. 4-10). At least for the period since 1871 with exception of winter and since 1850 for all seasons, the warming has been generally stronger in the north than south of 60° N over the Baltic Sea region.

A closer look on the phenomenon of polar amplification is provided by Fig. 4.10. Latitudinal differences in warming trends indeed increased the latitudinal temperature gradient over the Baltic Sea in all seasons with exception of winter in recent decades. This is shown by 51yr running north-south temperature gradients as anomalies from the long-term mean temperature gradient. The comparison of 51yr running decadal trends in the north and south shows a relatively synchronous variation for both regions in summer and spring with mostly higher magnitudes in the north. This was also the case for autumn till the 1960s when the southern region shows decreasing trends and did not follow the warming in the northern region. Most remarkable is the low and partly opposing co-variability of decadal trends in winter. While all 51yr. periods after the early 1950s show a warming trend in the south, a strong decadal cooling occurred in the north during the 1940s to 1960s. Only in recent decades did winter warming trends in the north supersede decadal warming trends in the southern region.

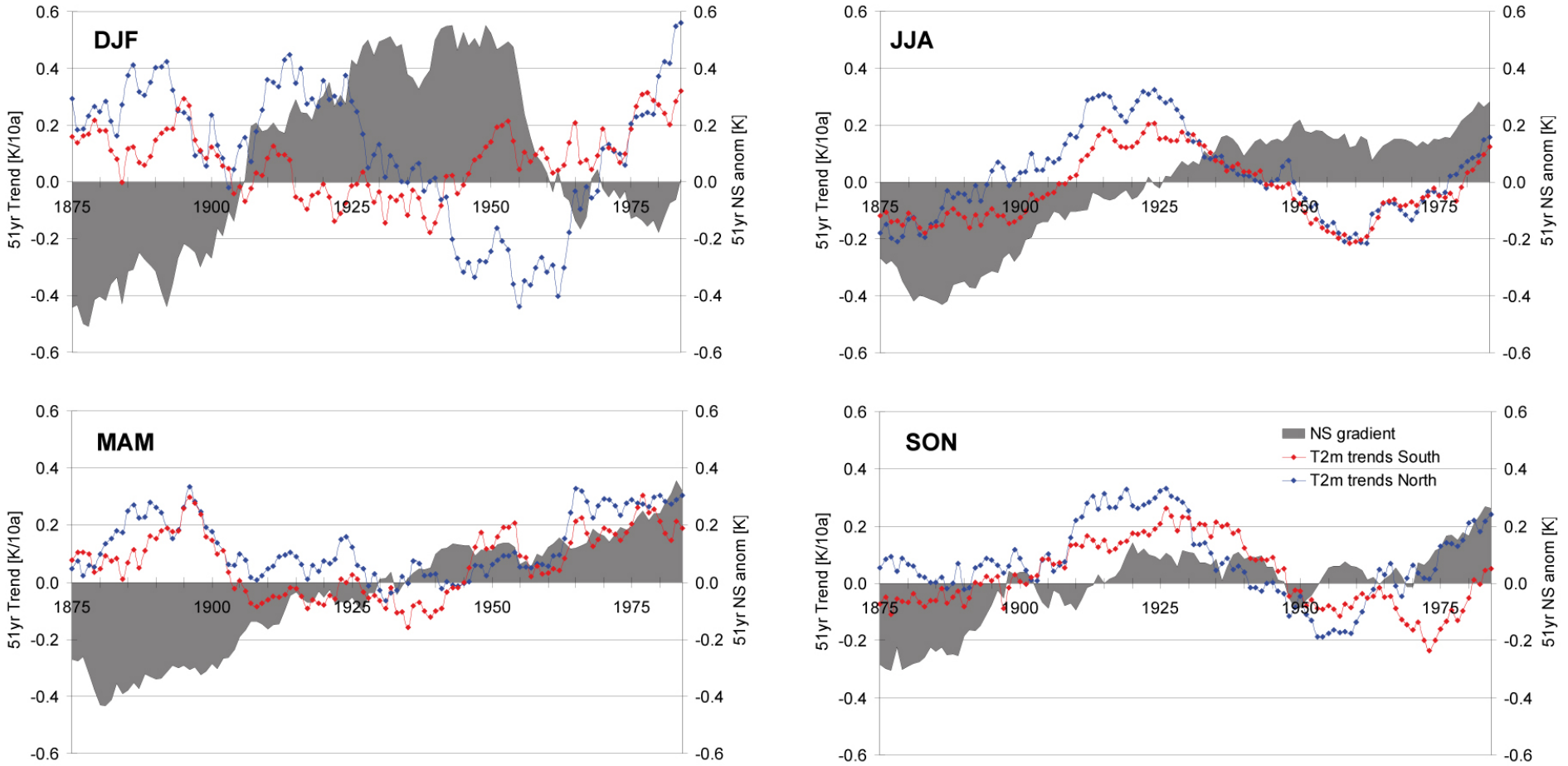

Fig. 4-10: Time series of 51yr running trends [K/10a] north (blue, 66°-70° N, 20°-30° E) and south (red, 48°-52° N, 20°-30° E) of the Baltic Sea and the corresponding anomalies [K] of 51yr running north-south gradients (gray) of seasonal mean temperatures for the period 1850-2009. Years correspond to the centre of 51yr periods.

4.5 Conclusion

The analysis in this chapter was applied to estimate the long-term temporal and spatial agreement of the reconstruction compared to gridded observations since 1901 and 1871 and a set of station observations since 1850. The question was here to which extent the temperature conditions are realistically reconstructed over the whole period of 160 years and to which extent potential deviations of the reconstruction may be dependent on model fields of RCAX.

The results regarding the correct reconstruction of seasonal and annual warming trends by analog-upscaling from 22 stations provide high confidence into the temperature evolution of HiResAFF since 1850. Regional warming trends are almost identically reconstructed compared to CRUTEM3 over the Baltic Sea (BACC 2008) since 1871 with exception of underestimated warming in spring. The latter demonstrates the dependency of analog-upscaling on adequate target fields by the used regional model. As the simulation of RCAX fails to reproduce very cold and warm conditions compared to observations since 1961, the relatively low sample size of analogs does not allow a large improvement by the reconstruction. The cross-comparison with the RCAX model, from which target fields have been re-sampled, shows that the magnitude of trends are generally better captured by the reconstruction while the model is better in reconstructing the regional pattern in comparison to gridded observations of CRU TS.

The model dependency of the reconstruction method mostly relates to spatial characteristics of the used model. On the one hand, this is exactly the purpose of analog-upscaling linking station observations to physically consistent fields from a regional climate model. On the other hand, the spatial deficiencies by the used model like negative pattern correlations and bias have to be accepted. In contrast, variations, long-term trends and spatial trend patterns are relatively independently reconstructed leading to a partly better agreement of HiResAFF with observations than with RCAX. An important result is here that regional changes e.g. of latitudinal temperature gradients and spatiotemporal different time evolutions of decadal trends can be successfully reconstructed.

Combined with the conclusions from chapter 2 (Schenk and Zorita, 2012), the analog reconstruction successfully reconstructs high-frequency (daily) temperature anomalies predicted by daily SLP and monthly mean temperatures predicted by monthly mean T2m observations. Although the variability on monthly and seasonal scale is underestimated between 5-20%, long-term variations and regional decadal to centennial warming trends are realistically reconstructed at least since 1850. The general disadvantage of reconstructing monthly mean temperature analogs from a considerable smaller analog pool (~150 analogs for each month) limited to a certain period (1961-2007) does not lead to evident problems to reconstruct seasonal variations and trends. Only the case of spring suggest that a higher sample size might reduce the deficiencies by RCAX over the re-sampling of the analog-method.

5 Applications – the Baltic Sea state since 1850

The reconstructed multivariate atmospheric forcing fields of HiResAFF have been used to drive three different ecosystem models of the Baltic Sea since 1850. Part of the project ECOSUPPORT (Meier et al., 2012a; http://www.baltex-research.eu/ecosupport/index.html) was aimed at reconstructing the transition of the Baltic Sea from its pristine state around 1850 through the strong increase of human induced nutrification since the 1950s until today. Continuing the historical simulation with future greenhouse gas emission (GHG) scenario simulations until 2100, possible future changes caused by climate change and/or eutrophication can be studied within the context and range of long-term changes since 1850. The ensemble of three different ecosystem models driven by the same meteorological fields of HiResAFF, river-runoff and nutrient loads provides the so far longest transient simulation of the Baltic Sea evolution and allows a simultaneous inter-model-comparison including a first quantification of model uncertainties (Meier et al., 2012b). The main results with relevance to the used forcing of HiResAFF are summarized in section 5.1. More details about the historical simulations have been published in Gustafsson et al. (2012) and Meier et al. (2012b).

As an important part of the modeling activities, numerous consistency tests have been applied to estimate whether HiResAFF allows a realistic simulation of key processes of the Baltic Sea. Of main interest are here high-frequency pressure and wind variations which dominate horizontal and vertical mixing and irregularly occurring inflow events of saline and oxygenized water from the North Sea. As one example, the reconstructed atmospheric conditions around the strongest inflow event to the Baltic Sea 1951/52 are compared with those of NCEP reanalysis in section 5.2. Results from a preliminary analysis about the frequency of similar atmospheric conditions and potential changes in related wind directions since 1850 are presented. Section 5.3 shortly discusses the consistency of the simulated long-term seasonal sea-surface-temperature variations and simulated sea-ice extent since 1850 compared to historical observations.

5.1 Evolution of the Baltic Sea state since 1850

With reconstructed meteorological forcing of HiResAFF, river-runoff and reconstructed nutrient loads, transient simulations by three different ecosystem models are used to reconstruct of the Baltic Sea state over the period 1850-2006 (Meier et al., 2012b, Gustafsson et al., 2012). The comparison with available observations suggests a realistic reconstruction of eutrophication from very low levels until the beginning of the 20th century with a strong increase since the 1950s. A key result of the long-term simulation is that the timescales of eutrophication are exceptionally long (Gustafsson et al., 2012). Although nutrient loads have been clearly reduced since 1980, the phosphorus content reached only recently an equilibrium with the loads. This implies that no improvements of the water quality have been achieved by drastically reducing nutrient loads by one third since 1980.

Although the timescales and involved processes need to be further studied, the long timescales of eutrophication become also visible in scenario simulations of the Baltic Sea state until 2100 (Meier et al., 2012b). Even optimistic nutrient load reduction scenarios lead to little or no improvement of the water quality in case of phosphate or oxygen (Fig. 5-1). Even in case of stabilized nutrient contents, the simulation forced by atmospheric conditions under A2 and A1B emission scenarios very likely leads to a further reduction of bottom oxygen and hence increasing amount of hypoxia (dead zones) (Fig. 5-1). The results suggest that stronger nutrient load reductions need to be considered by the HELCOM Baltic Sea action plan.

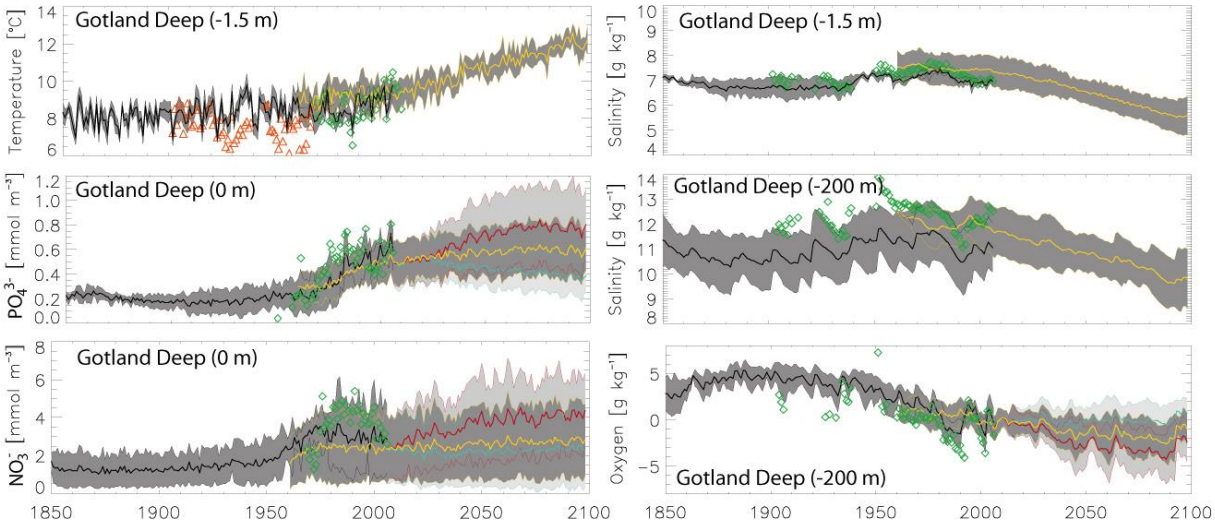

Fig. 5-1: Simulation results for the Gotland Deep from three ecosystem models: Ensemble mean and ±1-sigma uncertainty for annual mean water temperatures (1.5 m depth), January to March mean surface phosphate and nitrate concentration, annual mean salinity (1.5 and 200 m depth) and oxygen concentration (200 m depth). Meteorological forcing 1850-2006 by HiResAFF, 1961-2098 by RCM downscaled GHG scenarios (A2, A1B). For details see Meier et al. (2012b).

5.2 Meteorological conditions linked to inflow events

A prerequisite for an adequate reproduction of salinity and oxygen by the ecosystem models in Fig. 5-1 is a realistic meteorological forcing regarding high-frequency (at least daily) pressure and wind data with realistic variability, amplitudes and extremes. With exception of a systematic offset of bottom (200 m depth) salinity in the Gotland Deep, the simulations very closely reproduce the temporal evolution of salinity and oxygen relative to observations since the beginning of the 20th century (green diamonds, Fig. 5-1). Apart from the fresh water budget, which is controlled by river-runoff and evaporation, salinity and the ventilation of deep basins with bottom oxygen depends on inflow events from the North Sea.

The atmospheric conditions of such inflow events usually consist of two episodes. First, a longer period of easterly winds with high pressure is needed to decrease the sea-level of the Baltic Sea by pushing out brackish surface water to the North Sea. If this episode is followed by an abrupt turn to strong westerly winds, an inflow of saline and oxygenized water from the North Sea may reach the deep basins like the Gotland deep. A period of only strong westerly winds is in contrast not able to produce significant inflow events as the sea-level of the Baltic Sea allows no filling up of the basins. Figure 5-3 shows an example for the pressure and wind conditions related to the strongest inflow event to the Baltic Sea since observations started in 1881. The real conditions become even more complicated as salinity and water temperature gradients in the usually barotropic transition zone suppress the exchange of surface and bottom water through the Belts.

Based on these conditions, only 82 major inflows have been observed in the period 1897-1976 (Fischer and Matthäus, 1996) with almost no inflow events after the 1980s (Fig. 5-2). Model simulations from 1902-1998 (Meier and Kauker, 2003) produce 180 major inflow events while a semi-empirical model by Gustafsson and Andersson (2001) produced 118 events. The generally good agreement between simulated and measured salinity and bottom oxygen in Fig. 5-1 on long timescales suggests the HiResAFF and the models are able to reproduce these events. This increases also the confidence into the scenario simulations until 2100.

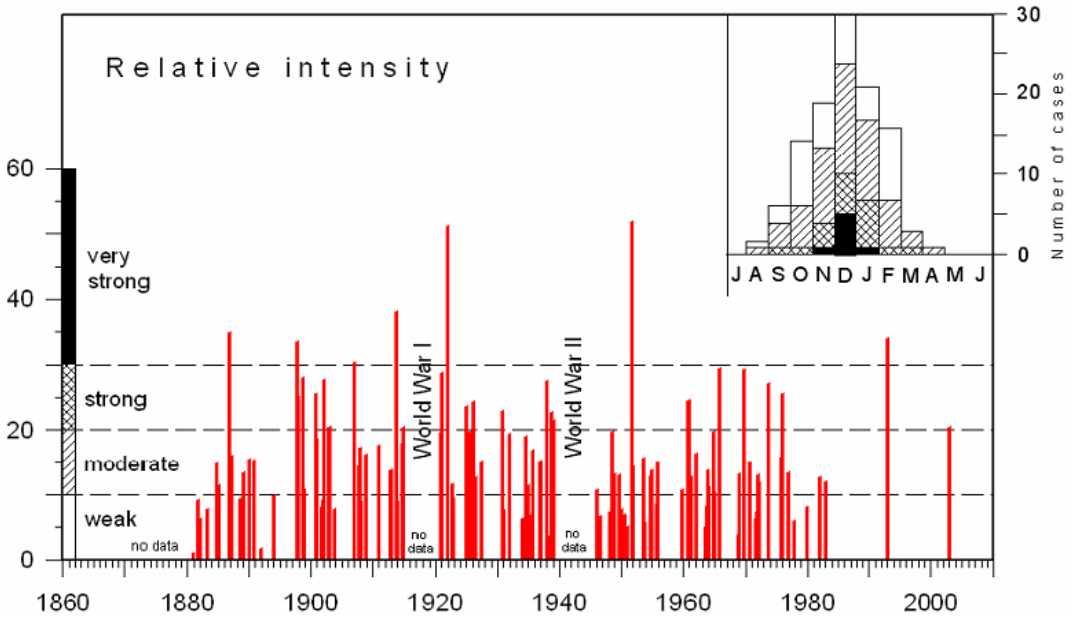

Fig. 5-2: Frequency and relative intensity of inflow events since 1881 (updated and reproduced by SMHI after Fischer and Matthäus, 1996).

An unprecedented absence of inflow events since the 1990s in Fig. 5-2 led to strongly increasing hypoxia and decreasing bottom salinity in the Baltic Sea which is also captured by the simulations in Fig. 5-1. While hypoxic areas have been already strongly increased by human induced eutrophication, a combination with unfavorable (natural or potentially changed future) weather conditions can lead to severe water problems. As the Baltic Sea currently holds the world's largest human induced dead zones (Conley, 2012), a brief analysis of pressure and wind conditions related to inflow events have been performed for HiResAFF.

The comparison of daily pressure and wind conditions around the inflow invent 1951/52 from HiResAFF (green) and NCEP (blue) (Fig. 5-3) yields a very good agreement both in phase and magnitudes. The red arrows mark the period of high pressure, a constant negative zonal (easterly) flow with low surface wind speeds from mostly easterly sectors. This period led to a significant outflow and low sea-levels of the Baltic Sea basin. The abrupt turn to very strong westerly winds from end of November 1951 and afterwards (blue arrows) allowed a very strong inflow of heavy saline and oxygenized water to the deep basin of the Baltic Proper.

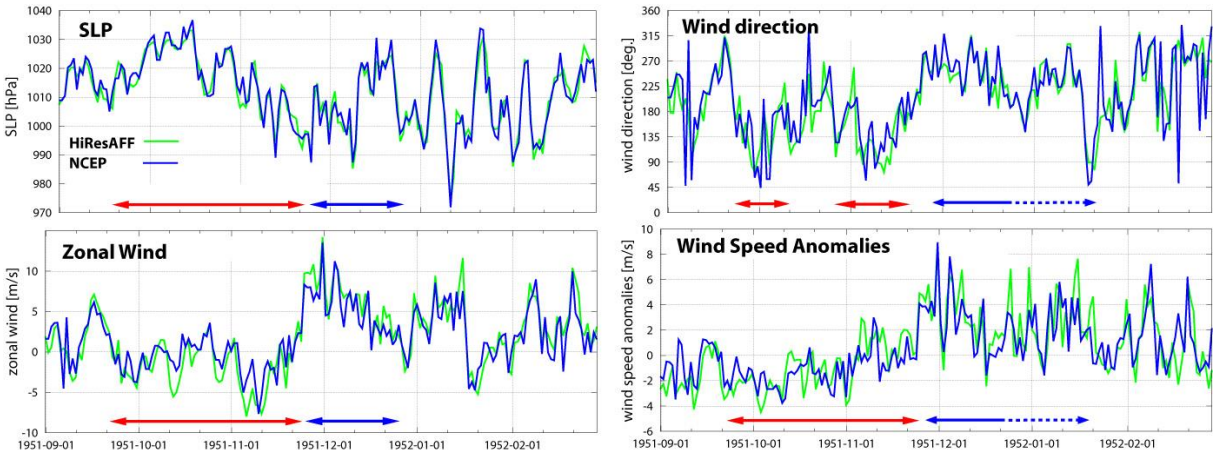

Fig. 5-3: Comparison of daily pressure and wind conditions over the western Baltic Sea area (52.5-60° N; 10-17.5° E) during the inflow event 1951/52 derived from HiResAFF and NCEP.

The absence of observed inflow events since the 1990s corresponds to the very stormy period and the strong NE-shift of deep lows discussed in chapter 4. As strong westerly winds lead to high sea-levels in the Baltic Sea, no inflow events are able to enter the Baltic Sea basin. To estimate whether any change in the frequency of wind directions might explain the situation at the end of the 20th century, frequencies of daily easterly wind directions over the Kattegat are calculated from HiResAFF for every month from September to April since 1850. As the exact wind directions might not be relevant, also NNE/NE winds (22.5-90°) are considered in addition to winds from NE (45-90°).

Figure 5-4 shows the relative frequency of north-easterly wind directions on monthly basis for the months October to March as anomalies to the monthly long-term mean frequency. All months show generally large annual and decadal variations. Consistent with the stormy period since the 1980s and the NE-shift of deep lows in winter and spring (Lehmann et al., 2011, chapter 4), easterly wind directions have clearly declined. I.e. January and March show a pronounced clustering of strong negative anomalies in the last decades.

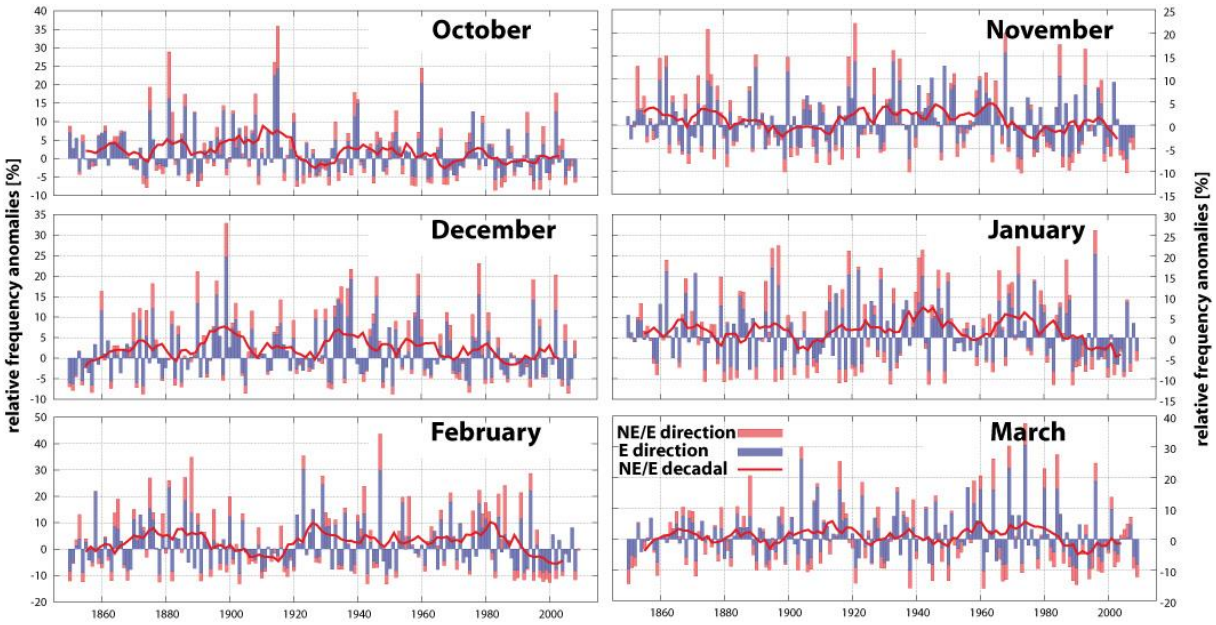

Fig. 5-4: Relative monthly frequency of daily NE/E (red) and E-wind (blue) directions over the western Baltic Sea (Kattegat) as anomalies to the long-term mean frequency. An 11yr running mean is used to highlight decadal variations of NE/E-wind directions. Note different scaling.

Although a more detailed analysis is required to study the relation of pressure and wind variations related to inflow conditions (cf. Zorita and Laine, 2000), the results indicate a clear link between anomalies of NE/E-wind directions with unprecedented low frequency inflow events since the 1980s (Fig. 5-2). To a less extent, these conditions are also found for the stormy 1870s when no inflow observations are available. Results from laminated sediment cores of the Baltic Sea suggest that periods of widespread hypoxia occurred throughout the Holocene during warm periods of the Medieval and Holocene Thermal Maximum (Zillén et al., 2008). Zillén and Conley (2010) even suggest that human influences have already contributed to hypoxia during the last two millennia. To which extent a future decrease of salinity and oxygen in the scenario simulations (Fig. 5-1) are also related to changes in the wind climate needs to be further analyzed. If the NE-shift and increasing impact of deep lows over Scandinavia since the 1970s (chapter 4) continues under climate change conditions, further deteriorations of the water quality are possible apart from the influence of increasing temperatures and nutrient loads.

5.3 Sea-surface-temperatures and ice-extent of the Baltic Sea

As shown in chapter 4, the reconstructed seasonal warming trends of HiResAFF over deep sea areas like the Gotland Deep are lower than air temperatures over land. As a consequence, simulated SSTs in Fig. 5-1 also show a slight warming trend from 1850 to 2006 which is in the range of measurements by the light ship Svenska Björn 1902-1968 (orange triangles) and monitoring cruises at Gotland Deep since 1970 (green diamonds). Only the period since 1980 shows a strong warming trend. A detailed validation of seasonal mean SSTs for different locations with light ship observations since 1860 shows very realistic results and stationary reconstruction skills for the BALTSEM simulation (r~0.7) (see Gustafsson et al., 2012).

For winter conditions, the annual maximum sea-ice extent of the Baltic Sea (MIB) can be considered as an integrated value over the area averaged time evolution of air temperatures over the Baltic Sea from around November to February or March. A comparison of the simulated MIB by the ecosystem of BALTSEM (Gustafsson et al., 2012) with observations (updated by FMI after Seinä and Palosuo, 1996) shows a good agreement on annual (r=0.72) and decadal (r=0.84) scale (Fig. 5-5).

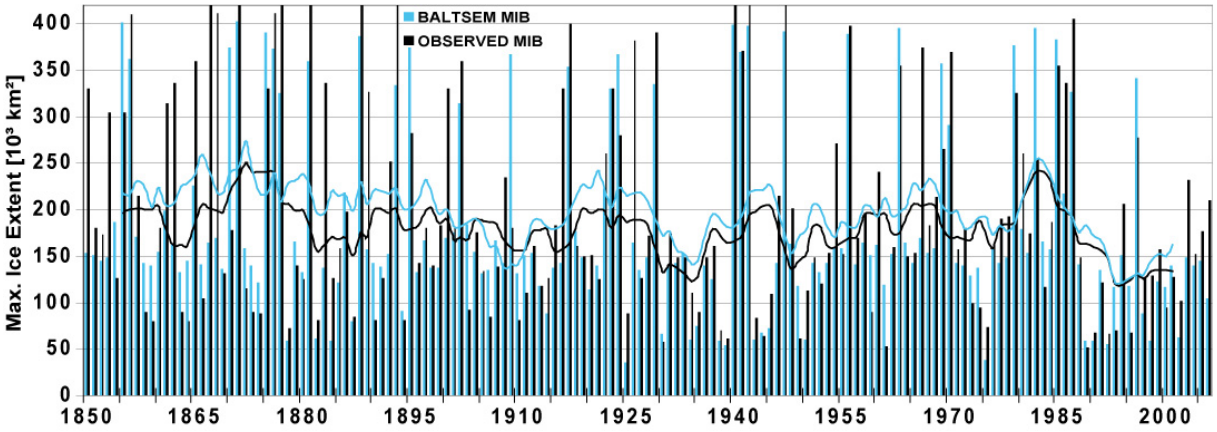

Fig. 5-5: Comparison of the observed annual maximum ice-extent of the Baltic Sea (MIB) [10^3 km²] with the simulation of BALTSEM forced by HiResAFF 1850-2006 (r=0.72). The lines show the filtered 11yr running mean of the MIB.

As highlighted by the running decadal mean MIB, the simulated values [10^3 km²] are systematically higher (mean=198.3, median=161, sigma=110.2) than the observations (mean=178.4, median=148, sigma=98.3). This might be related to the model and forcing bias and/or the coarse resolution of BALTSEM. The latter becomes evident in a scatter plot between simulated and observed MIB values with partly large absolute annual deviations for very high/low values (not shown). The scatter plot of ranked MIB values shows in comparison a more regular scattering along the regression line with r=0.72.

Both datasets show a very similar long-term trend from 1850 to 2006 with a clear decrease of 37.000 km² (-18.8%, observations) or 34.000 km² (-19.2%, BALTSEM) per century. This trend is mainly caused by high ice-extents around 1870 and very low values since 1990. Apart from this episodes, the time series is rather dominated by large annual to decadal scale variations with no gradual change in between. Recent years not captured by the figure demonstrated the dominance of large variations with a very high MIB value of 309.000 km² in 2011/12 and the record low value of 49.000 km² in 2008/09. Nevertheless, the unprecedented low ice-cover at the end of the 20th century indicates again that the NE-shift of deep lows towards Scandinavia (chapter 4) plays a more important role for the regional climate of the Baltic Sea than centennial long-term trends in air temperatures.

6 Summary and Conclusion

The aim of this thesis was to develop and apply a new approach for the reconstruction of physically consistent atmospheric fields from sparse historical station records. As the required properties of these fields are ideally provided by regional climate models (RCM), the analog-method is introduced as a statistical upscaling tool which links a set of homogenized long-term station records to analogous fields from the RCM. In chapter 2, different settings for the analog-upscaling have been evaluated to optimize the reconstruction and tests its robustness. In chapter 3, reconstructed long-term variations and trends in annual storminess were analysed and compared with other studies over Northern Europe since 1850 followed by a comparison of reconstructed long-term variations and trends of seasonal temperatures in chapter 4. Brief evaluations of Baltic Sea ecosystem model simulations driven by the reconstructed forcing fields confirm a realistic reconstruction both for high- and low-frequency variations (chapter 5). In the following, the main results regarding the principal concept of analog-upscaling, the dependency on the used RCM and the robustness of the long-term application for wind, temperature and ecosystem modelling are summarized.

6.1 Characteristics of analog-upscaling

In the first part of the thesis, the analog-method was introduced, evaluated and applied as statistical upscaling tool to reconstruct physically consistent meteorological fields suitable to drive ocean-, ecosystem or similar models. Based on the maximized pattern similarity between a target day and its analog in station records, the method assumes that corresponding fields of the analog should also be very similar to those of the target day with unknown atmospheric fields. The analog-reconstruction is hence completed by reconstructing the target fields in the past by the fields of the closest analog provided by the RCM in recent decades.

By design, analog-upscaling realistically reproduces daily variance and variable- and region-specific frequency distributions, spatial patterns etc. by re-sampling the RCM fields in time. As these properties can be also achieved by randomly resampled fields, the benchmark of the reconstruction is to achieve good temporal correlations through the predictive skill of station records. As demonstrated in chapter 2, the assumption about analogous fields and hence good temporal correlations is valid if the atmospheric fields show a close physical link to the predictor variable. Taking daily station pressure as predictor, this is the case for pressure, wind, and with seasonal differences also precipitation and temperature fields. For relative humidity and cloudiness, correlations on daily and monthly scale are on average rather low. Using block-bootstrapping, it is shown that the amount of locally significant correlations at the 5% level is sufficiently high so that field significance can be claimed for all variables in comparison to the original RCM simulation.

Several tests have been applied to test the robustness of the reconstruction skill (correlations) dependent on the number and spatial distributions of predictors, the number of analogs, the stationary of the calibration period or the density of suitable next neighbours to the best analog (chapter 2). The analysis shows that little or no improvement is achieved when daily analogs for a 50 year test reconstruction are taken from an archive of 10 years ($N_{Analogs}$~900 per target month) or 50 years ($N_{Analogs}$~4500 per month). Also the calibration period, e.g. by taking different 10 year junks for the 50 year reconstruction, has a negligible effect. Of higher importance for the correlation i.e. on daily scale seems to be the used RCM fields which is in principal independent from the availability of suitable analogs within the predictor time series.

6.2 Dependency on the used RCM fields and spectral nudging

In the archive of analogs, the station observations of a given day are linked to the fields of the same date of the RCM. Because the RCM is driven by reanalysis data ("observations") at its lateral boundaries, the fields are expected to be closely related to the predictor stations. This link is a prerequisite for the assumption that analogs within the station records have also analogous atmospheric fields. The comparison of a reconstruction performed within the surrogate climate of the RCM and the reconstruction using real station pressure as predictor revealed a strong gap in the level of correlations on daily and to a less extent on monthly scale (chapter 2). In the surrogate approach, the clearly higher correlations show the optimal skill of analog-upscaling as the used grid points used as pseudo-predictors are consistent with the target fields of the model. In contrast, the lower correlations found when using real station data as predictor can be explained by the fact that the observations are only partly correlated with the reanalysis driven fields of the RCM. As the reanalysis drives the RCM only at the boundaries, the simulation in the interior of the domain can evolve more freely and at least partly independent from observations.

One possibility suggested in chapter 2 is to use spectral nudging which brings the RCM closer to observations also in the interior of the domain. Based on the same analogs, but using the fields taken from a spectrally nudged RCM, the reconstruction has been recently repeated, displaying clearly higher daily correlations than in the case of the reconstruction without spectral nudging (not shown). The improvement becomes less relevant on monthly scales so that little effects can be expected for the long-term skill. The evaluation of the reconstructed temperatures since 1850 included in chapter 4 also suggested a dependency on deficiencies of the RCM, affecting the reconstructed long-term trends and spatial patterns when they are compared to observations. While spatial patterns of gridded observations might also be not fully consistent, further tests need to be applied to ascertain to what extent these properties can be further improved. As spatial patterns might change over time beyond those in the analog-pool, analog-upscaling cannot be expected to capture them all. One possibility to get more independent samples from an analog-pool could be to search analogs for a set of station records/proxies in the space of RCM fields e.g. of paleoclimate simulations.

6.3 Robustness of analog-upscaling on longer timescales

In the context of long-term reconstructions, a very important aspect relates to the dependency of correlations on the number and spatial distribution of predictors (e.g. proxies). The example of reconstructing wind fields with artificially reduced numbers of pressure stations as predictors shows mainly an effect also visible with higher number of stations. Regions and seasons which are strongly linked to large-scale pressure variations like sea areas of NW-Europe in winter need only low numbers of predictors while more continental regimes and summer conditions need a considerably higher amount of predictors. This will generally also depend on the variables of the predictors and the target fields. These regional- and variable-specific dependency of the reconstruction skill on the number of predictors should be considered e.g. when using different numbers/locations of predictors. A first test case in chapter 3 shows that differences in long-term trends of storminess caused by reducing the number of predictors are not negligible in regions away from the predictors. Hence, changes in the number/location of predictors might introduce spurious results over time in remote regions while this effect is rather small if at least some predictors are present in a specific region. The chosen strategy to use only those predictors which cover at least more than the 20th century rather than incorporating all available records over time gives a high confidence into the long-term robustness of the reconstruction.

6.3.1 Consistency of long-term storminess over Northern Europe

The analysis and comparison of reconstructed long-term variations and trends of annual storminess with the 20th Century Reanalysis (20CR) and different pressure-based storm indices suggests that the reconstructed storminess is consistent with pressure-based storm indices derived from homogenized station pressure. This is the case for geostrophic storm indices over the NE-Atlantic and North Sea and single-station indices of Stockholm since 1850. Dependent on the index, fair to very high correlations exist on annual and decadal scale with consistent long-term trends. For geostrophic indices derived from not homogenized pressure information, the reconstruction and 20CR suggest that some periods of these indices might be problematic as e.g. secondary stations had to be used for a limited time period. Nevertheless, these indices, the reconstruction and a thorough literature review about long-term trends in the Euro-Atlantic storm climate lead to the conclusion that no robust long-term trends but rather large (multi-)decadal variations dominated the last 200 to 300 years. Only the analysis of the spatial area affected by deep cyclones over Scandinavia has indicated that unprecedented (since 1850) high values have been reached at the end of the 20th century with a clear impact on regional storminess. Unprecedented low maximum ice extents of the Baltic Sea (chapter 5) have been also documented since that period. While also externally forced AOGCM simulations of the last millennium do not suggest any link between low-frequency temperature variations and storminess, it needs to be further analysed whether the spatial shift of storm tracks observed and reconstructed since the 1970s is part of large multi-decadal variations or due to external forcing.

The comparison of reconstructed fields with 20CR confirms an earlier study indicating large inconsistencies in 20CR's long-term storminess trends over the NE-Atlantic and North Sea in comparison to observations. By comparing decadal scale correlations of detrended and original storm time series of 20CR with the reconstruction, the differences suggest that the inconsistencies of spurious long-term trends in 20CR are large over open sea areas but considerably smaller over land areas of Northern Europe. As these discrepancies disappear after around 1940, the reconstruction casts further doubt about the suitability of 20CR for the analysis of long-term trends i.e. over data sparse regions back in time. As the reconstruction keeps the number of predictors relatively constant at least since 1874, the high agreement with other historical storm observation proof the reliability of the pressure-based analog-reconstruction also on longer timescales.

6.3.2 Reproduction of long-term temperature variations and trends

As the mean value and long-term variations and trends of temperature have only a weak physical link to pressure, an alternative strategy for analog-upscaling was introduced in chapter 2. While daily temperature anomalies are reasonably reproduced by daily pressure as predictor, the monthly means have been predicted by monthly mean station records. Although this approach yields a good reconstruction skill in the test period of 50 years in chapter 2, a principal disadvantage of reconstructing monthly analogs relates to the considerably reduced number of available analogs ($N_{Analogs}$~150 per target month). Chapter 4 demonstrated nevertheless that the reconstruction reproduces the long-term seasonal variations and trends reasonably well. Regional differences in trends and hence spatial patterns are rather related to the bias caused by the RCM than by the reconstruction method.

The better agreement of long-term trends on longer timescales with gridded observations of CRUTEM3 since 1871 indicates that the longer period to resample the fields can partly reduce deviations introduced by the original RCM simulation. Nevertheless, a larger analog pool should be considered when reconstructing monthly analogs and/or temperature changes

on longer timescales. With an increasing length of the reconstruction, the analog-pool might less likely contain suitable analogs for considerable different climates in the past if it evolves outside the observational range. This might be relevant e.g. in the context of paleoclimate reconstructions. Here, the approach previously suggested that analogs for the predictors/proxies are directly search in the RCM fields e.g. of a paleoclimate simulation, might be a useful further application of analog-upscaling. Also the introduction of scaling might help to overcome the principal disadvantage of the analog approach in comparison to linear regression methods.

6.4 Applications

The results of three ecosystem simulations of the Baltic Sea since 1850 prove that the reconstructed forcing fields through analog-upscaling provide realistic conditions both for high- and low-frequency variations. The realistic reconstruction of high-frequency variations of pressure and wind is reflected by a generally good agreement of simulated oxygen and salinity concentrations with observations. This includes a realistic reproduction of irregularly occurring inflow events from the North Sea which are primarily determined by high-frequency variations and extremes of wind and pressure conditions. A brief analysis of monthly variations in NE/E wind directions since 1850 suggests that the unprecedented absence of inflow events since the 1990s is related to clearly reduced frequencies of north-easterly winds. This period corresponds the period of unprecedented high impacts of deep lows.

A detailed validation of simulated SSTs by the ecosystem model BALTSEM with historical measurements since 1860 show high annual correlations which are also robust over different periods (for details see online material to Gustafsson et al., 2012). The good agreement of simulated and observed annual and decadal variations of the maximum sea-ice extent of the Baltic Sea since 1850 further confirms a robust skill over the whole period. The very similar observed and simulated decrease of the ice extent by about ~19% per century since 1850 is mainly a result of very low values after 1980 and very high values around 1870 with little gradual change in between.

Bibliography

Alexander, L.V.; Tett, S.F.B. & Jonsson, T. (2005): Recent observed changes in severe storms over the United Kingdom and Iceland. *Geophys. Res. Lett.*, 32, L13704, doi:10.1029/2005GL022371

Alexandersson, H.; Schmith, T.; Iden, K. & Tuomenvirta, H. (1998): Long-term variations of the storm climate over NW Europe. *Global Atmos. Ocean. Syst.*, 6, 97-120.

Alexandersson, H.; Tuomenvirta, H.; Schmith, T. & Iden, K. (2000): Trends of storms in NW Europe derived from an updated pressure data set. *Clim. Res.*, 14, 71-73, doi:10.3354/cr014071

Ansell, T. J., Jones, P.D., Allan, R.J. et al. (2006): Daily mean sea level pressure reconstructions for the European - North Atlantic region for the period 1850-2003. *J. Climate*, 19, 2717-2742, doi:10.1175/JCLI3775.1

Auer, I, Böhm, R, Jurkovic, A. et al. (2007): HISTALP – Historical instrumental climatological surface time series of the greater Alpine region 1760-2003. *Int. J. Climatol.*, 27, 17-46, doi:10.1002/joc.1377

BACC (2008): Assessment of Climate Change for the Baltic Sea Basin. Regional Climate Studies. Springer, Berlin/Heidelberg.

Barnett, T. and Preisendorfer, R. (1978): Multifield analog prediction of short-term climate fluctuations using a climate state vector. *J. Atmos. Sci.*, 35, 1771–1787, doi:10.1175/1520-0469(1978)035<1771:MAPOST>2.0.CO;2

Bärring, L. & von Storch, H. (2004): Scandinavian storminess since about 1800. *Geophys. Res. Lett.*, 31, L20202, doi:10.1029/2004GL020441

Bärring, L. & Fortuniak, K. (2009): Multi-indices analysis of southern Scandinavian storminess 1780-2005 and links to interdecadal variations in the NW Europe-North Sea region. *Int. J. Climatol.*, 29, 373-384, doi:10.1002/joc.1842

Barriopedro, D.; García-Herrera, R.; Lupo, A. R. & Hernández, E. (2006): A Climatology of Northern Hemisphere Blocking J. Climate, Journal of Climate, *American Meteorological Society*, 19, 1042-1063, doi:10.1175/JCLI3678.1

Bengtsson, L., Hagemann, S. & K.I. Hodges (2004): Can climate trends be calculated from reanalysis data? *J. Geophys. Res.*, 109, D11111, doi:10.1029/2004JD004536

Bhend, J., and von Storch, H. (2008): Consistency of observed winter precipitation trends in northern Europe with regional climate change projections. *Clim. Dynam.*, 31, 17-28, doi:10.1007/s00382-007-0335-9

Bhend, J., and von Storch, H. (2009): Is greenhouse gas forcing a plausible explanation for the observed warming in the Baltic Sea catchment area? *Boreal Environ. Res.*, 14, 81-88.

Biau, G., Zorita, E., von Storch, H., and Wackernagel, H. (1999): Estimation of precipitation by kriging in the EOF space of the sea levelpressure field. *J. Climate*, 12, 1070–1085, doi:10.1175/1520-0442(1999)012<1070:EOPBKI>2.0.CO;2

Böhm, R.; Jones, P.; Hiebl, J.; Frank, D.; Brunetti, M. & Maugeri, M. (2010): The early instrumental warm-bias: a solution for long central European temperature series 1760-2007. *Climatic Change*, 101, 41-67, doi:10.1007/s10584-009-9649-4

Brohan, P.; Kennedy, J. J.; Harris, I.; Tett, S. F. B. & Jones, P. D. (2006): Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850 *J. Geophys. Res.*, 111, D12106, doi:10.1029/2005JD006548

Bromwich, D.H., Fogt, R.L., Hodges, K.I. & J.E. Walsh (2007): A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. *J. Geophys, Res.*, 112, D10111, doi:10.1029/2006JD007859

Brönnimann, S.; Martius, O.; von Waldow, H.; Welker, C.; Luterbacher, J.; Compo, G.; Sardeshmukh, P. & Usbeck, T. (2012): Extreme winds at northern mid-latitudes since 1871. *Meteorol. Z.*, 21, 13 - 27, doi:10.1127/0941-2948/2012/0337

Brunet, M. and Jones, P. (2011): Data rescue initiatives: bringing historical climate data into the 21st century. *Clim. Res.*, 47, 29–40, doi:10.3354/cr00960

Bürger, G., Fast, I., and Cubasch, U. (2006): Climate reconstruction by regression - 32 variations on the theme. *Tellus*, 58A, 227-235, doi:10.1111/j.1600-0870.2006.00164.x

Carter, D. J. T. & Draper, L. (1988): Has the north-east Atlantic become rougher? *Nature*, 332, 494-494, doi:10.1038/332494a0

Chang, E. K. M. & Fu, Y. (2002): Interdecadal Variations in Northern Hemisphere Winter Storm Track Intensity. *J. Climate*, 15, 642-658, doi:10.1175/1520-0442(2002)015<0642:IVINHW>2.0.CO;2

Cheung, K.Y., and Lee, S.M.S (2005): Variance estimation for sample quantiles using the m out of n bootstrap. *Ann. Inst. Statist. Math.*, 57, 279-290, doi:10.1007/BF02507026

Christensen, J., Kjellström, E., Giorgi, F., Lenderink, G., and Rummukainen, M. (2010): Weight assignment in regional climate models. *Clim. Res.*, 44, 179–194, doi:10.3354/cr00916

Ciavola, P.; Ferreira, O.; Haerens, P.; Van Koningsveld, M.; Armaroli, C. & Lequeux, Q. (2011): Storm impacts along European coastlines. Part 1: The joint effort of the MICORE and ConHaz Projects. *Environmental Science & Policy*, 14, 912-923, doi:10.1016/j.envsci.2011.05.011

Conley, D. J. (2012): Ecology: Save the Baltic Sea. Nature, 486, 463-464, doi:10.1038/486463a

Compo, G. P.; Whitaker, J. S.; Sardeshmukh, P. D. et al. (2011): The Twentieth Century Reanalysis Project. *Q. J. R. Meteorol. Soc.*, 137, 1-28, doi:10.1002/qj.776

Cubasch, U., von Storch, H., Waszkewitz, J., and Zorita, E. (1996): Estimates of climate changes in southern Europe using different downscaling techniques, *Clim. Res.*, 7, 129–149, doi:10.3354/cr007129

Dangendorf, S.; Mudersbach, C.; Jensen, J.; Anette, G. & Heinrich, H. (2013): Seasonal to decadal forcing of high water level percentiles in the German Bight throughout the last century. *Ocean Dynamics*, 1-16, doi:10.1007/s10236-013-0614-4

Dee, D. P., Uppala, S. M., Simmons, A. J. et al. (2011): The ERA Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. Roy. Meteor. Soc.*, 137, 553–597, doi:10.1002/qj.828

de Kraker, A. M. J. (1999): A method to assess the impact of high tides, storms and storm surges as vital elements in climatic history. The case of stormy weather and dikes in the northern part of Flanders, 1488 to 1609. *Climatic Change*, 43, 287-302, doi:10.1023/A%3A1005598317787

Déqué, M., Jones, R.G., Wild, M., Giorgi, F., Christensen, J.H., Hassell, D.C., Vidale, P.L., Rockel, B., Jacob, D., Kjellström, E., de Castro, M., Kucharski, F., and van den Hurk, B. (2005): Global high resolution versus Limited Area Model climate change projections over Europe: quantifying confidence level from PRUDENCE results, *Clim. Dynam*., 25(6), 653-670, doi:10.1007/s00382-005-0052-1

Donat, M. G.; Renggli, D.; Wild, S.; Alexander, L. V.; Leckebusch, G. C. & Ulbrich, U. (2011): Reanalysis suggests long-term upward trends in European storminess since 1871. *Geophys. Res. Lett.*, 38, L14703, doi:10.1029/2011GL047995

Döscher, R., Willén, U., Jones, C., Rutgersson, A., Meier, H.E.M., Hansson, U., and Graham, L.P. (2002): The development of the regional coupled ocean-atmosphere model RCAO, *Boreal Environ. Res.*, 7, 183–192.

Ebisuzaki, W. (1997): A method to estimate the statistical significance of correlation when the data are serially correlated. *J. Climate*, 10, 2147-2153, doi:10.1175/1520-0442(1997)010<2147:AMTETS>2.0.CO;2

Efron, B. (1982): The jackknife, the Bootstrap and other resampling plans. J.W. Arrowsmith Ltd., Bristol, England, doi:10.1137/1.9781611970319

Feser, F., Barcikowska, M., Krueger, O., Schenk, F., Weisse, R., and L. Xia (2014): Storminess over the North Atlantic and Northwestern Europe - A Review. QJRMS, doi:10.1002/qj.2364

Ferguson, C. R. and G. Villarini (2012): Detecting inhomogeneities in the Twentieth Century Reanalysis over the central United States. *J. Geophys. Res.*, 117 (D05123), doi:10.1029/2011JD016988

Fernández, J. and Sáenz, J. (2003): Improved field reconstruction with the analog method: searching the CCA space. *Clim. Res.*, 24, 199–213, doi:10.3354/cr024199

Fischer, H. & Matthäus, W. (1996): The importance of the Drogden Sill in the Sound for major Baltic inflows. *Journal of Marine Systems*, 9, 137-157, doi:10.1016/S0924-7963(96)00046-2

Fischer-Bruns, I.; von Storch, H.; González-Rouco, J. F. & Zorita, E. 2005): Modelling the variability of midlatitude storm activity on decadal to century time scales. *Clim. Dynam.*, 25, 461-476, doi:10.1007/s00382-005-0036-1

Frank, D.; Büntgen, U.; Böhm, R.; Maugeri, M. & Esper, J. (2007): Warmer early instrumental measurements versus colder reconstructed temperatures: shooting at a moving target. *Quaternary Science Reviews*, 26, 3298 - 3310, doi:10.1016/j.quascirev.2007.08.002

Frías, D., Zorita, E., Fernández, J., and Rodríguez-Puebla, C. (2006): Testing statistical downscaling methods in simulated climates. *Geophys. Res. Lett.*, 33, L19807, doi:10.1029/2006GL027453

Giorgi, F., Bi, X., and Pal, J. (2004): Means, trends and interannual variability in a regional climate change experiment over Europe. Part I: Present day climate (1961-1990). *Clim. Dynam.*, 22(6-7), 733-756, doi:10.1007/s00382-004-0409-x

Graham, N.E., Hughes, M.K., Ammann, C.M., Cobb, K.M., Hoerling, M.P., Kennett, D.J., Kennett, J.P., Rein, B., Stott, L., Wigand, P.E., and Xu, T. (2007): Tropical Pacific - Mid-latitude Teleconnections in Medieval Times. *Climatic Change*, 83, 241-285, doi:10.1007/s10584-007-9239-2

Graham, L. P., Olsson, J., Kjellström, E., Rosberg, J., Hellstöm, S.-S., and Berndtsson, R. (2009): Simulating river flow to the Baltic Sea from climate simulations over the past millennium, *Boreal Env. Res.*, 14: 173–182.

Guiot, J., Corona, C., and ESCARSEL members (2010): Growing Season Temperatures in Europe and Climate Forcings Over the Past 1400 Years. *PLoS ONE* 5(4): e9972, doi:10.1371/journal.pone.0009972

Gulev S. K.; Zolina, O. & Grigoriev, S. (2001): Extratropical cyclone variability in the Northern Hemisphere winter from the NCEP/NCAR reanalysis data. *Clim. Dynam.*, 17, 795-809, doi:10.1007/s003820000145

Gustafsson, B. G. & Andersson, H. C. (2001): Modeling the exchange of the Baltic Sea from the meridional atmospheric pressure difference across the North Sea. *J. Geophys. Res.*, 106, 19731-19744, doi:10.1029/2000JC000593

Gustafsson B.G., Schenk F., Blenckner T., Eilola K., Meier H.E.M., Müller-Karulis B., Neumann T., Ruoho-Airola T., Savchuk O.P. & Zorita, E. (2012): Reconstructing the Development of Baltic Sea Eutrophication 1850–2006. *AMBIO*, 41, 534-548, doi:10.1007/s13280-012-0318-x

Hanna, E.; Cappelen, J.; Allan, R.; Jónsson, T.; Le Blancq, F.; Lillington, T. & Hickey, K. (2008): New Insights into North European and North Atlantic Surface Pressure Variability, Storminess, and Related Climatic Change since 1830. *J. Climate*, 21, 6739-6766, doi:10.1175/2008JCLI2296.1

Harris, I., Jones, P.D., Osborn, T.J., and Lister, D.H. (2013): Updated high-resolution grids of monthly climatic observations. *Int. J. Climatol.*, 34: 623–642, doi:10.1002/joc.3711

HELCOM (2011): Fifth Baltic Sea Pollution Load Compilation (PLC-5). Baltic Sea Environmental Proceedings 128. Helsinki, Finland.

Hickey, K. R. (2003): The storminess record from Armagh Observatory, Northern Ireland, 1796-1999. *Weather*, 58, 28-35, doi:10.1256/wea.293.01

Hines, K.M., Bromwich, D.H. & G.J. Marshall (2000): Artificial Surface Pressure Trends in the NCEP–NCAR Reanalysis over the Southern Ocean and Antarctica. *J. of Climate*, 13: 3940-3952, doi:10.1175/1520-0442(2000)013%3C3940:ASPTIT%3E2.0.CO;2

Hogben, N. (1995): Increases in wave heights over the North Atlantic: A review of the evidence and some implications for the naval architect. *Trans. Roy. Inst. Naval Arch.*, W5, 93-101.

Hurrell, J.W. (1995): Decadal trends in the North Atlantic Oscillation and relationships to regional temperature and precipitation. *Science*, 269, 676-679, doi:10.1126/science.269.5224.676

Hurrell, J.W. and C.K. Folland (2002): A change in the summer atmospheric circulation over the North Atlantic. *CLIVAR Exchanges Newsletter*, 7, 52-54.

Hurrell, J.; Kushir Y.; Ottersen G. & Visbeck, M. (2003): The North Atlantic Oscillation. Climatic Significance and Environmental Impact. *Geophysical Monograph*, 134, 279pp., doi:10.1029/GM134

Jacobeit, J., H. Wanner, J. Luterbacher, C. Beck, A. Philipp and K. Sturm (2003): Atmospheric circulation variability in the North Atlantic-European area since the mid-seventeenth century. *Clim. Dyn.*, 20, 341-352, doi:10.1007/s00382-002-0278-0

Jones, P. D. (1994): Hemispheric Surface Air Temperature Variations: A Reanalysis and an Update to 1993. *J. Climate*, 7, 1794-1802, doi:10.1175/1520-0442(1994)007<1794:HSATVA>2.0.CO;2

Jones, P.D., and Moberg, A. (2003): Hemispheric and large-scale surface air temperature variations: An extensive revision and an update to 2001. *J. Climate*, 16, 206-223, doi:10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2

Jones, P. D.; Lister, D. H.; Osborn, T. J.; Harpham, C.; Salmon, M. & Morice, C. P. (2012): Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Geophys. Res.*, 117, D05127, doi:10.1029/2011JD017139

Jun, M., Knutti, R., and Nychka, D.W. (2008): Spatial Analysis to Quantify Numerical Model Bias and Dependence. *J. A. Stat. Assoc.*, 103, 934-947, doi:10.1198/016214507000001265

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woolen, J., Chelliah, M., Ebiszusaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., and Fiorino, M. (2001): The NCEP-NCAR 50 year reanalysis. *B. Am. Meteorol. Soc.*, 82, 247–267, doi:10.1175/1520-0477(2001)082%3C0247:TNNYRM%3E2.3.CO;2

Kjellström, E., Nikulin, G., Hansson, U., Strandberg, G., and Ullerstig, A. (2011): 21st century changes in the European climate: uncertainties derived from an ensemble of regional climate model simulations. *Tellus*, 63A, 24-40, doi:10.1111/j.1600-0870.2010.00475.x

Klein Tank, A.M.G. et al. (2002): Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.*, 22, 1441-1453, doi:10.1002/joc.773

Krueger, O.; Schenk, F.; Feser, F. & Weisse, R. (2013): Inconsistencies between Long-Term Trends in Storminess Derived from the 20CR Reanalysis and Observations. *J. Climate*, 26, 868-874, doi:10.1175/JCLI-D-12-00309.1

Krueger, O. & von Storch, H. (2011): Evaluation of an Air Pressure–Based Proxy for Storm Activity. *J. Climate*, 24, 2612-2619, doi:10.1175/2011JCLI3913.1

Krueger, O. & von Storch, H. (2012): The informational value of pressure-based single-station proxies for storm activity. *J. Atmos. Oceanic Technol.*, 29, 569-580, doi:10.1175/JTECH-D-11-00163.1

Kruizinga, S. and Murphy, A.H. (1983): Use of an analogue procedure to formulate objective probabilistic temperature forecasts in the Netherlands. *Mon. Weather Rev.*, 111, 2244-2254, doi:10.1175/1520-0493(1983)111<2244:UOAAPT>2.0.CO;2

Lehmann, A.; Getzlaff, K. & Harlaß, J. (2011): Detailed assessment of climate variability in the Baltic Sea area for the period 1958 to 2009. *Clim. Res.*, 46, 185-196, doi:10.3354/cr00876

Lindenberg, J.; Mengelkamp, H.-T. & Rosenhagen, G. (2012): Representativity of near surface wind measurements from coastal stations at the German Bight. *Meteorol. Z.*, 21, 099-106, doi:10.1127/0941-2948/2012/0131

Liu, R.Y., Singh, K. (1992): Moving blocks bootstrap captures weak dependence. In: Exploring the Limits of the Bootstrap, Whiley, 225-248, 1992.

Livezey, R. E. and Chen, W. Y. (1983): Statistical field significance and its determination by Monte Carlo techniques. *Mon. Weather Rev.*, 111, 46–59, doi:10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2

Lorenz, E. N. (1969): Atmospheric predictability as revealed by naturally occurring analogs. *J. Atmos. Sci.*, 26, 639–646, doi:10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2

Luo, D. & Wan, H. (2005): Decadal variability of wintertime North Atlantic and Pacific blockings: A possible cause. *Geophys. Res. Lett.*, 32, L23810, doi:10.1029/2005GL024329

Matulla, C., Haas, P., Wagner, S., Zorita, E., Formayer, H., and Kromp-Kolb, H. (2004): Anwendung der Analog-Methode in komplexem Terrain: Klimaänderungsszenarien auf Tagesbasis für Österreich. GKSS Report 2004/11, 2004.

Matulla, C. (2005): Regional, seasonal and predictor-optimized downscaling to provide groups of local scale scenarios in the complex structured terrain of Austria. *Meteorol. Z.*, 14, 31-45, doi:10.1127/0941-2948/2005/0014-0031

Matulla, C.; Schöner, W.; Alexandersson, H.; von Storch, H. & Wang, X. (2007): European storminess: late nineteenth century to present. *Clim. Dynam.*, 31, 125-130, doi:10.1007/s00382-007-0333-y

Matulla, C., Zhang, X., Wang, X.L., Wang, J., Zorita, E., Wagner, S., and von Storch, H. (2008): Influence of similarity measures on performance of downscaling precipitation by the analog method for downscaling precipitation. *Clim. Dynam.*, 30 (2-3), 133-144, doi:10.1007/s00382-007-0277-2

McCabe, G.J., Clark, M.P., and M.C. Serreze (2001): Trends in Northern Hemisphere Surface Cyclone Frequency and Intensity. *J. of Climate*, 14: 2763-2768, doi:10.1175/1520-0442(2001)014%3C2763:TINHSC%3E2.0.CO;2

Meier, H. E. M. & Kauker, F. (2003): Modeling decadal variability of the Baltic Sea: 2. Role of freshwater inflow and large-scale atmospheric circulation for salinity. *J. Geophys. Res.*, 108, 3368, doi:10.1029/2003JC001799

Meier, H.E.M., Eilola, K., and Almroth, E. (2011a): Climate-related changes in marine ecosystems simulated with a 3-dimensional coupled physical-biogeochemical model of the Baltic Sea. *Clim. Res.*, 48, 31-55, doi:10.3354/cr00968

Meier, H.E.M., Höglund, A., Döscher, R., Andersson, H., Löption, U., and Kjellström, E. (2011b): Quality assessment of atmospheric surface fields over the Baltic Sea from an ensemble of regional climate model simulations with respect to ocean dynamics. *Oceanologa*, 53, 193–227.

Meier, H.E.M., and Andersson, H. (2012a): ECOSUPPORT: A Pilot Study on Decision Support for Baltic Sea Environmental Management. *AMBIO*, 41, 529–533, doi:10.1007/s13280-012-0317-y

Meier, H. E. M., Andersson, H. C., Arheimer, B. et al. (2012b): Comparing reconstructed past variations and future projections of the Baltic Sea ecosystem – first results from multi-model ensemble simulations. *Environ. Res. Lett.*, 7, 034005, doi:10.1088/1748-9326/7/3/034005

Meier, H. E. M., Andersson, H. C., Arheimer, B., Donnelly, C., Eilola, K., Gustafsson, B. G., Kotwicki, L., Neset, T.-S., Niiranen, S., Piwowarczyk, J., Savchuk, O. P., Schenk, F., Węsławski, J.-M. & E. Zorita (2014): Ensemble Modeling of the Baltic Sea Ecosystem to Provide Scenarios for Management. *AMBIO*, 43, 37-48, doi:10.1007/s13280-013-0475-6

Mitchell, T. D. & Jones, P. D. (2005): An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int. J. Climatol.*, 25, 693-712, doi:10.1002/joc.1181

Moberg, A.; Alexandersson, H.; Bergström, H. & Jones, P. D. (2003): Were southern Swedish summer temperatures before 1860 as warm as measured? *Int. J. Climatol.*, 23, 1495-1521, doi:10.1002/joc.945

Moberg, A., Sonechkin, D., Holmgren, K., Datsenko, N., and Karlen, W. (2005): Highly variable northern hemisphere temperatures reconstructed from low- and high resolution proxy data. *Nature*, 433, 613–617, doi:10.1038/nature03265

Omstedt, A. & Chen, D. (2001): Influence of atmospheric circulation on the maximum ice extent in the Baltic Sea. *J. Geophys. Res.*, 106, 4493-4500, doi:10.1029/1999JC000173

Omstedt, A., C. Pettersen, J. Rodhe and P. Winsor (2004): Baltic Sea climate: 200 yr of data on air temperature, sea level variations, ice cover and atmospheric circulation. *Clim. Res.*, 25, 205-216, doi:10.3354/cr025205

Parker, D. & Horton, B. (2005): Uncertainties in central England temperature 1878--2003 and some improvements to the maximum and minimum series. *Int. J. Climatol.*, 25, 1173-1188, doi:10.1002/joc.1190

Parker, D. E.; Legg, T. P. & Folland, C. K. (1992): A new daily central England temperature series, 1772-1991. *Int. J. Climatol.*, 12, 317-342, doi:10.1002/joc.3370120402

Pinto, J. G.; Zacharias, S.; Fink, A. H.; Leckebusch, G. C. & Ulbrich, U. (2009): Factors contributing to the development of extreme North Atlantic cyclones and their relationship with the NAO. *Clim. Dynam.*, 32, 711-737, doi:10.1007/s00382-008-0396-4

Polyakov, I. V.; Alekseev, G. V.; Bekryaev, R. V.; Bhatt, U.; Colony, R. L.; Johnson, M. A.; Karklin, V. P.; Makshtas, A. P.; Walsh, D. & Yulin, A. V. (2002): Observationally based assessment of polar amplification of global warming. *Geophys. Res. Lett.*, 29, 25-1-25-4, doi:10.1029/2001GL011111

Raible C.; Della-Marta P. M.; Schwierz C. & R., B. (2008): Northern Hemisphere extratropical cyclones: a comparison of detection and tracking methods and different reanalyses. *Mon. Weather Rev.*, 136, 880-897, doi:10.1175/2007MWR2143.1

Rimbu, N. and G. Lohmann (2011): Winter and summer blocking variability in the North Atlantic region – evidence from long-term observational and proxy data from southwestern Greenland. *Clim. Past*, 7, 543-555, doi:10.5194/cp-7-543-2011

Rockel, B., and Woth, K. (2007): Extremes of near-surface wind speed over Europe and their future changes as estimated from an ensemble of RCM simulations. *Climatic Change*, 81, *Supplement 1*, 267-280, doi:10.1007/s10584-006-9227-y

Rosenhagen, G., and Bork, I. (2009): Rekonstruktion der Sturmflutwetterlage vom 13. November 1872. *Die Küste*, 75, 51-70.

Samuelsson, P., Jones, C.G., Willén, U., Ullerstig, A., Gollvik, S., Hansson, U., Jansson, C., Kjellström, E., Nikolin, G., and Wyser, K. (2011): The Rossby Centre Regional Climate model RCA3: model description and performance. *Tellus*, 63A: 4-23, doi:10.1111/j.1600-0870.2010.00478.x

Schenk, F. and Zorita, E. (2012): Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling. *Clim. Past*, 8, 1681-1703, doi:10.5194/cp-8-1681-2012

Schiesser, H. H.; Pfister, C. & Bader, J. (1997): Winter storms in Switzerland North of the Alps 1864/1865-1993/1994. *Theor. Appl. Climatol.*, 58, 1-19, doi:10.1007/BF00867428

Schimanke, S., Meier, H. E. M., Kjellström, E., Strandberg, G., and Hordoir, R. (2012): The climate in the Baltic Sea region during the last millennium simulated with a regional climate model. *Clim. Past*, 8, 1419–1433, doi:10.5194/cp-8-1419-2012

Schmidt, H. & von Storch, H. (1993): German Bight storms analysed. *Nature*, 365, 791-791, doi:10.1038/365791a0

Seinä, A. & E. Palosuo (1996): The Classification of the Maximum Annual Extent of Ice Cover in the Baltic Sea 1720-1995 - Based on the material collected by Rirto Jurva (winters 1720-1940) and the material of the Ice Service of the Finnish Institute of Marine Research (winters 1941-1995). *Meri-Report Series of the Finnish Institute of Marine Research*, 27, 1996.

Serreze, M. C. & Barry, R. G. (2011): Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change*, 77, 85-96, doi:10.1016/j.gloplacha.2011.03.004

Sickmöller M.; Blender R. & Frädrich, K. (2000): Observed winter cyclone tracks in the northern hemisphere in re-analysed ECMWF data. *Quart. J. Roy. Meteorol. Soc.*, 126, 591-620, doi:10.1002/qj.49712656311

Suvilampi, E. (2009): Voimakkaiden geostrofisten tuulten alueellisuus ja muutokset Suomessa vuosina 1884-2100. Pro Gradututkielma. Turun yliopisto, maantieteen laitos, 2009.

Sweeney, J. (2000): A three-century storm climatology for Dublin 1715-2000. *Irish Geography*, 33, 1-14, doi:10.1080/00750770009478595

Thorne, P. W. & Vose, R. S. (2010): Reanalyses Suitable for Characterizing Long-Term Trends. Are They Really Achievable? *Bull. Amer. Meteor. Soc.*, 91, 353-361, doi:10.1175/2009BAMS2858.1

Toth, Z. (1989): Long-Range Weather Forecasting Using an Analog Approach. *J. Climate*, 2, 594–607, doi:10.1175/1520-0442(1989)002<0594:LRWFUA>2.0.CO;2

Trigo, I. F. (2006): Climatology and interannual variability of storm-tracks in the Euro-Atlantic sector: a comparison between ERA-40 and NCEP/NCAR reanalyses. *Clim. Dynam.*, 26, 127-143, doi:10.1007/s00382-005-0065-9

Trouet, V., Esper, J., Graham, N.E., Baker, A., Scourse, J., and Frank, D. (2009): Persistent positive North Atlantic Oscillation dominated the Medieval Climate Anomaly. *Science*, 324, 78-80. doi:10.1126/science.1166349

Ulbrich, U.; Leckebusch, G. C. & Pinto, J. G. (2009): Extra-tropical cyclones in the present and future climate: a review. *Theor. Appl. Climatol.*, 96, 117-131, doi:10.1007/s00704-008-0083-8

Uppala, S.M., Kållberg, P.W., Simmons, A.J., et al. (2006): The ERA-40 analysis. *Quart. J. Roy. Meteorol. Soc.*, 131, 2961-3012, doi:10.1256/qj.04.176

van den Dool, H. (1994): Searching for analogs, how long must we wait? *Tellus*, 46A, 314–324, doi:10.1034/j.1600-0870.1994.t01-2-00006.x

Vautard, R. and Yiou, P. (2009): Control of recent European surface climate change by atmospheric flow. *Geophys. Res. Lett.*, 36, L22702, doi:10.1029/2009GL040480

Vidale, P.L., Lüthi, D., Frei, C., Seneviratne, S., and Schär, C. (2003): Predictability and uncertainty in a regional climate model. *J. Geophys. Res.*, 108(D18), 4586, doi:10.1029/2002JD002810

von Storch, H., Zorita, E., and Cubasch, U. (1993): Downscaling of global climate change estimates to regional scales: an application to Iberian rainfall in wintertime. *J. Climate*, 6, 1161-1171, doi:10.1175/1520-0442(1993)006<1161:DOGCCE>2.0.CO;2

von Storch, H., Zorita, E., Jones, J.M., Dimitriev, Y., Gonzalez-Rouco, F., and Tett, S. (2004): Reconstructing past climate from noisy data. *Science*, 306: 679–682, doi:10.1126/science.1096109

von Storch, H., and Zwiers, F. (1998): Statistical Analysis in Climate Research. Cambridge Univ. Press, New York, USA, 484pp., 1998.

Wang X. L.; Swail V. R. & Zwiers, F. W. (2006): Climatology and changes of extratropical cyclone activity: comparison of ERA40 with NCEP-NCAR reanalysis for 1958-2001. *J. Climate*, 19, 3145-3166, doi:10.1175/JCLI3781.1

Wang, X.; Zwiers, F.; Swail, V. & Feng, Y. (2009a): Trends and variability of storminess in the Northeast Atlantic region, 1874-2007. *Clim. Dyn.*, 33, 1179-1195, doi:10.1007/s00382-008-0504-5

Wang, X.; Swail, V.; Zwiers, F.; Zhang, X. & Feng, Y. (2009b): Detection of external influence on trends of atmospheric storminess and northern oceans wave heights. *Clim. Dyn.*, 32, 189-203, doi:10.1007/s00382-008-0442-2

Wang, X. L.; Feng, Y.; Compo, G. P.; Swail, V. R.; Zwiers, F. W.; Allan, R. J. & Sardeshmukh, P. D. (2012): Trends and low frequency variability of extra-tropical cyclone activity in the ensemble of Twentieth Century Reanalysis. *Clim. Dyn.*, 40, 2775-2800, doi:10.1007/s00382-012-1450-9

Wanner, H., Brönnimann, S., Casty, C., Gyalistras, D., Luterbacher, J. Schmutz, C. Stephenson, D. B., and Xoplaki, E. (2001): North Atlantic Oscillation - Concepts and Studies. *Surveys in Geophysics*, 22: 321–382, doi:10.1023/A%3A1014217317898

WASA group (1998): Changing Waves and Storms in the Northeast Atlantic? *Bull. Amer. Meteor. Soc.*, 79, 741-760, doi:10.1175/1520-0477(1998)079<0741:CWASIT>2.0.CO;2

Wetterhall, F., Halldin, S., and Xu, C. (2005): Statistical precipitation downscaling in central Sweden with the analogue method. *J. Hydrol.*, 306, 174-190, doi:10.1016/j.jhydrol.2004.09.008

Woodworth, P. L. & Blackman, D. L. (2002): Changes in extreme high waters at Liverpool since 1768. *Int. J. Climatol.*, 22, 697-714, doi:10.1002/joc.761

Xia, L.; von Storch, H. & Feser, F. (2012): Quasi-stationarity of centennial Northern Hemisphere midlatitude winter storm tracks. *Clim. Dyn.*, doi:10.1007/s00382-012-1543-5

Yoshimura, K. and Kanamitsu, M. (2008): Dynamical Global Downscaling of Global Reanalysis. *Mon. Weather Rev.*, 136, 2983–2998, doi:10.1175/2008MWR2281.1

Zillén, L., Conley, D. J., Andrén, T., Andrén, E., and Björck, S. (2008): Past occurrences of hypoxia in the Baltic Sea and the role of climate variability, environmental change and human impact. *Earth Sci. Rev.*, 91, 77–92, doi:10.1016/j.earscirev.2008.10.001

Zillén, L. and Conley, D.J. (2010): Hypoxia and cyanobacteria blooms - are they really natural features of the late Holocene history of the Baltic Sea? *Biogeosciences*, 7, 2567–2580, doi:10.5194/bg-7-2567-2010

Zorita, E. & Laine, A. (2000): Dependence of salinity and oxygen concentrations in the Baltic Sea on large-scale atmospheric circulation. *Clim. Res., Inter-research*, 14, 25-41, doi:10.3354/cr014025

Zorita, E., Hughes, J., Lettenmaier, D., and von Storch, H. (1995): Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *J. Climate*, 8, 1023-1042, doi:10.1175/1520-0442(1995)008<1023:SCORCP>2.0.CO;2

Zorita, E., and von Storch, H. (1999): The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *J. Climate*, 12, 2474-2489, doi:10.1175/1520-0442(1999)012%3C2474:TAMAAS%3E2.0.CO;2

List of Publications

Chapter 2 of this thesis has been published as Schenk and Zorita (2012) and parts of chapter 5 in Gustafsson et al. (2012). Minor contributions from this thesis were published in Meier et al. (2012b, 2014) and a more detailed discussion of the literature review than presented in chapter 3 was published in the review paper by Feser et al. (2014).

- Schenk, F. and Zorita, E. (2012): Reconstruction of high resolution atmospheric fields for Northern Europe using analog-upscaling. *Clim. Past*, 8, 1681-1703, doi: 10.5194/cp-8-1681-2012.

- Gustafsson B.G., Schenk F., Blenckner T., Eilola K., Meier H.E.M., Müller-Karulis B., Neumann T., Ruoho-Airola T., Savchuk O.P. & E. Zorita (2012): Reconstructing the Development of Baltic Sea Eutrophication 1850–2006. *AMBIO*, 41, 534-548, doi: 10.1007/s13280-012-0318-x.

- Meier, H. E. M., Andersson, H. C., Arheimer, B., Blenckner, T., Chubarenko, B., Donnelly, C., Eilola, K., Gustafsson, B. G., Hansson, A., Havenhand, J., Höglund, A., Kuznetsov, I., MacKenzie, B. R., Müller-Karulis, B., Neumann, T., Niiranen, S., Piwowarczyk, J., Raudsepp, U., Reckermann, M., Ruoho-Airola, T., Savchuk, O. P., Schenk, F., Schimanke, S., Väli, G., Węsławski, J.-M. & E. Zorita (2012b): Comparing reconstructed past variations and future projections of the Baltic Sea ecosystem - first results from multi-model ensemble simulations. *Environ. Res. Lett.*, 7, 034005, doi:10.1088/1748-9326/7/3/034005.

- Feser, F., Barcikowska, M., Krueger, O., Schenk, F., Weisse, R., & L. Xia (2014): Storminess over the North Atlantic and Northwestern Europe - A Review. *Quart. J. Roy. Meteorol. Soc.*, doi: 10.1002/qj.2364

- Meier, H. E. M., Andersson, H. C., Arheimer, B., Donnelly, C., Eilola, K., Gustafsson, B. G., Kotwicki, L., Neset, T.-S., Niiranen, S., Piwowarczyk, J., Savchuk, O. P., Schenk, F., Węsławski, J.-M. & E. Zorita (2014): Ensemble Modeling of the Baltic Sea Ecosystem to Provide Scenarios for Management. *AMBIO*, 43, 37-48, doi: 10.1007/s13280-013-0475-6

Acknowledgements

First of all, I want to thank Hans von Storch and Eduardo Zorita for supervising my thesis, for their motivation, support and sometimes patience and i.e. for many good hints and fruitful discussions. This includes also Sebastian Wagner who gave always scientific, technical and personal support whenever needed. I'm grateful for their motivation to rethink state-of-the-art concepts and try new alternatives. I want also to thank Oliver Krüger with whom I had many lively discussions in the last year and Christoph Matulla (ZAMG Vienna) for discussing his previous work using analog-downscaling. The work would not have been possible without the various support of many colleagues from different groups at the institute of HZG.

I also want to thank Markus Meier from SMHI Norrköping for managing the project ECOSUPPORT and providing the atmospheric fields from the regional climate model and Bo Gustafsson from the Baltic Nest Institute in Stockholm for fruitful discussions about the usability of the reconstruction as forcing for the ecosystem modelling of the Baltic Sea. I would like to thank also Lars Bärring (SMHI), Tuija Ruoho-Airola (FMI, Helsinki), Ari Venäläinen (FMI), Christine Luge (University of Jena), Gerard van der Schrier (KNMI) and Gudrun Rosenhagen (DWD Hamburg) for their help to provide and update station data.

Finally I'm very grateful to my family and friends who supported and motivated me over the four years of the thesis and who did not complain about not having me seen very often in the last stage of the thesis.

The research leading to these results has partly received funding from the European Union Seventh Framework Programme (FP/2007-2013) under grant agreement no. 217246 made with the joint Baltic Sea research and development programme BONUS, and the German Federal Ministry of Education and Research (03F0492A).

Erklärung selbstständige Ausarbeitung / Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den 16. Mai 2013