Clim. Past, 8, 1681–1703, 2012
www.clim-past.net/8/1681/2012/
doi:10.5194/cp-8-1681-2012

Climate of the Past

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

F. Schenk[1,2] **and E. Zorita**[1,2]

[1]Helmholtz-Zentrum Geesthacht, Institute for Coastal Research, Geesthacht, Germany
[2]KlimaCampus, University of Hamburg, Hamburg, Germany

Correspondence to: F. Schenk (frederik.schenk@hzg.de)

Received: 21 February 2012 – Published in Clim. Past Discuss.: 12 March 2012
Revised: 25 July 2012 – Accepted: 30 September 2012 – Published: 26 October 2012

Abstract. 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.

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, 2012; Gustafsson et al., 2012).

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). This idea is illustrated in Fig. 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

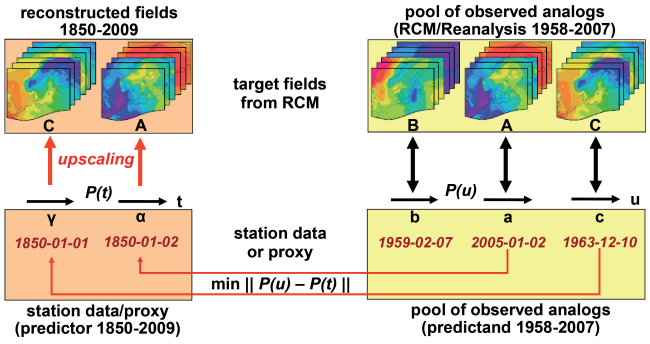

Fig. 1. Scheme for the Analog-Method used as upscaling tool. Any day a, b, c, etc. of $P(u)$ is linked to its related fields A, B, C, etc. taken from RCM/Reanalysis (predictand, analogs). The fields of a historical day γ, α, etc. from $P(t)$ is found by Eq. (1) to be most similar to c, a, etc. in $P(u)$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

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 (Meier et al., 2011a, 2012; 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.

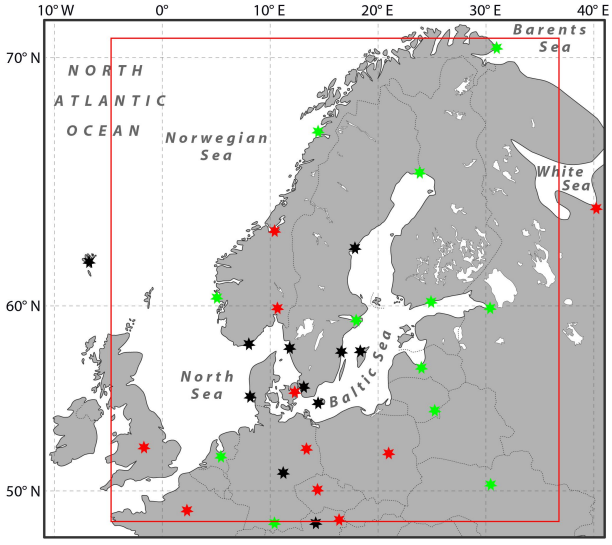

Fig. 2. Geographic positions of stations used as predictor in this study. Green stations provide daily SLP and monthly T2M, black stations only daily SLP and red stations only monthly T2M. The domain of the reconstruction is indicated by the red rectangle.

2 Data and methods

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) 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 compared to neighbouring stations.

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. 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 forcings, 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. 2.3.4). 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). 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 Atmospheric fields (predictand, analogs)

Focusing on Northern Europe and the Baltic Sea region, forcing fields with high spatio–temporal resolution are required in order to capture the high complexity of Baltic Sea subbasins. 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 with daily resolution.

2.3 Methods

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. 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:

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

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. 2.3.4).

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. 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. 3.2.1. Different settings are, however, required for the reconstruction of T2M (see below).

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. 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 $(n\text{lag} + 1) \cdot 23$. 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. 3.3.5).

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. 3.3.5, Fig. 10).

Monthly mean temperature fields are reconstructed separately using 22 stations providing monthly mean as predictor (Fig. 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.

2.4 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. 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 50 yr covered by the reference data using standard settings. The retrieved correlations for the different test cases are shown in Sect. 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. 3.2.2). The mean field correlation of the different tests are depicted in Fig. 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. 5, Sect. 3.2.3).

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. 2.3.4). 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. 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 of RCAX.

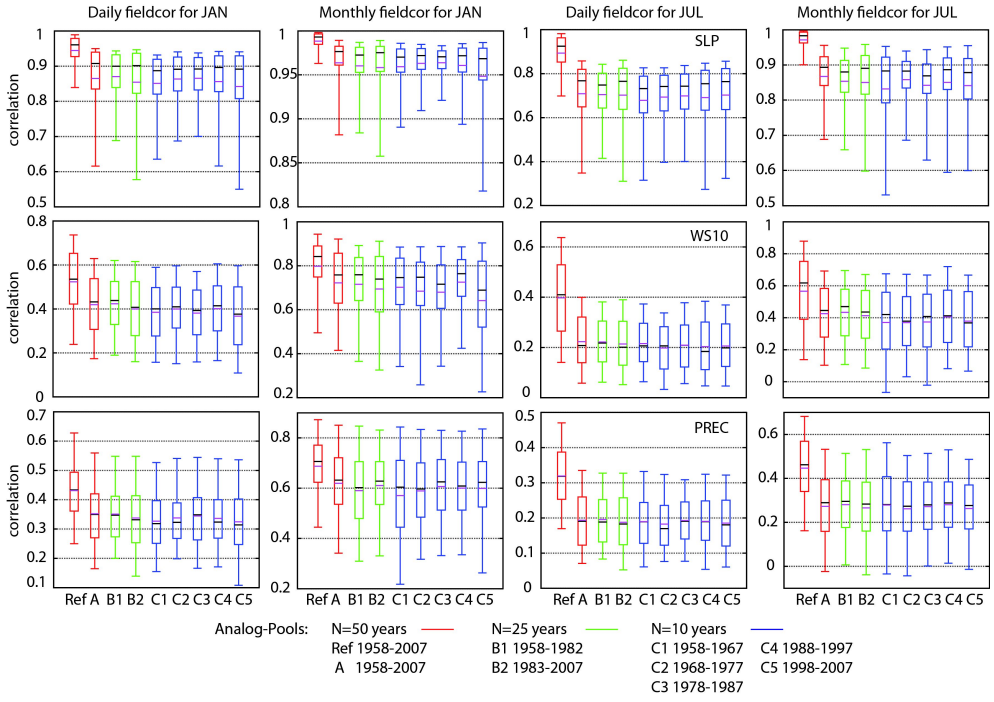

Fig. 3. Whisker-Box-Plots showing the field correlations of different test cases based on reconstructions from analog-pools of different size and periods. The box indicates the range of local correlations between the first and third quartile representing 50 % of the local correlations around the median (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.

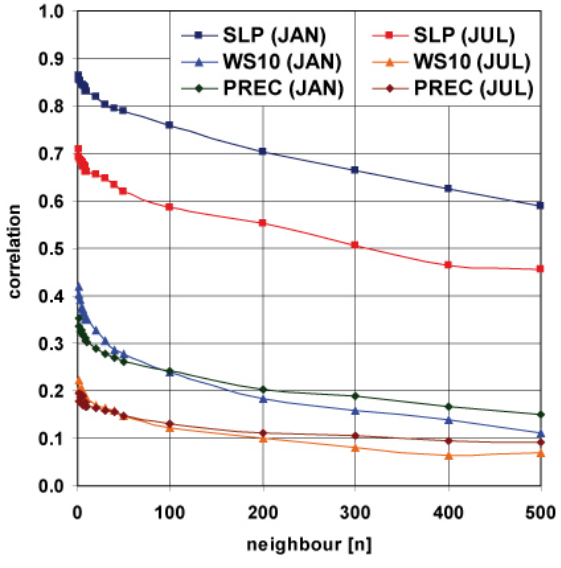

Fig. 4. Density of suitable analogs for HiResAFF estimated as the decay in daily mean field correlation as a function of n next neighbours instead of the best analog chosen from around 4500 possible analogs. Displayed are the variables of SLP, wind speed (WS) and PREC for January and July.

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 $\phi = \mathrm{var(REC)/var(REF)}$ 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.

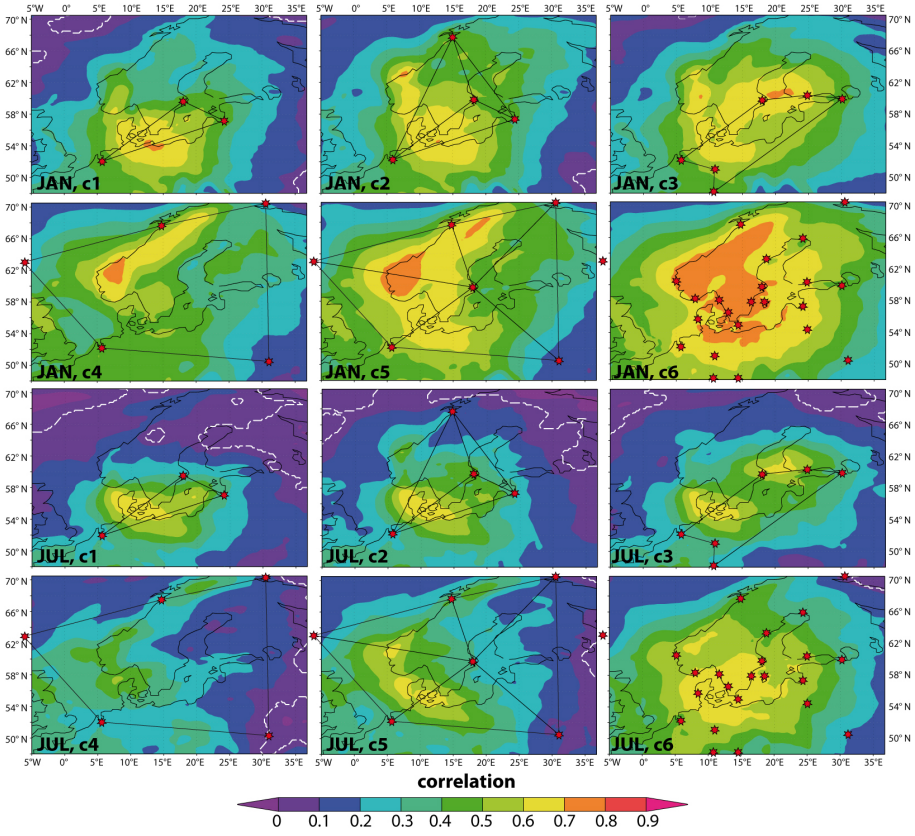

Fig. 5. Correlation of daily wind speeds between a surrogate reconstruction and reference fields of RCAO (RCAX for T2M) for January and July 1958–2007 dependent on the number and distribution of SLP predictors. White shaded lines indicate areas where h_0 of zero correlation cannot be rejected with $p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

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 95th quantile of the distribution of N sets is the 95th significance level. If the number N_{real} determined from the actual reconstructions is higher than the 95th 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 1–3, respectively.

3 Results

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 $r^2(\mathrm{Ref}) - r^2(\mathrm{A})$ 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).

3.2 Robustness of the analog-method

3.2.1 Dependency on 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 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).

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.

Table 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 %
RELHUM	0.23	97.8 %	0.13	81.3 %	0.46	76.2 %	0.24	40.3 %
TCLOUD	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 %

Table 2. Mean ratio of variance 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)	
	Rv(d)	h [%]	Rv(d)	h [%]	Rv(m)	h [%]	rv(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 %
RELHUM	1.05	*15.2 %*	1.20	57.3 %	0.58	54.1 %	0.39	81.2 %
TCLOUD	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 %*

Table 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	
	Δm	h [%]	Δm	h [%]
SLP	*0.04* hPa	*0.0* %	0.34 hPa	82.8 %
WS10	-0.20 m s^{-1}	*0.0* %	-0.32 m s^{-1}	68.0 %
PREC	-2.57 mm	*1.5* %	-8.23 mm	44.4 %
RELHUM	0.35 %	29.0 %	-0.37 %	49.8 %
TCLOUD	-0.30 %	10.3 %	-1.20 %	51.5 %
T2M	0.17 K	20.2 %	-0.03 K	3.3 %

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),

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.

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.

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. 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. 2.4, 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 in January, 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.

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 between HiResAFF and the reference fields $\mathrm{rv} = \phi = \sigma_{\mathrm{HiResAFF}}/\sigma_{\mathrm{RCAO}}$. Figure 7 shows the ratio of variance ϕ 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.

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 ($\phi < 0.5$), i.e. over the SE domain.

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

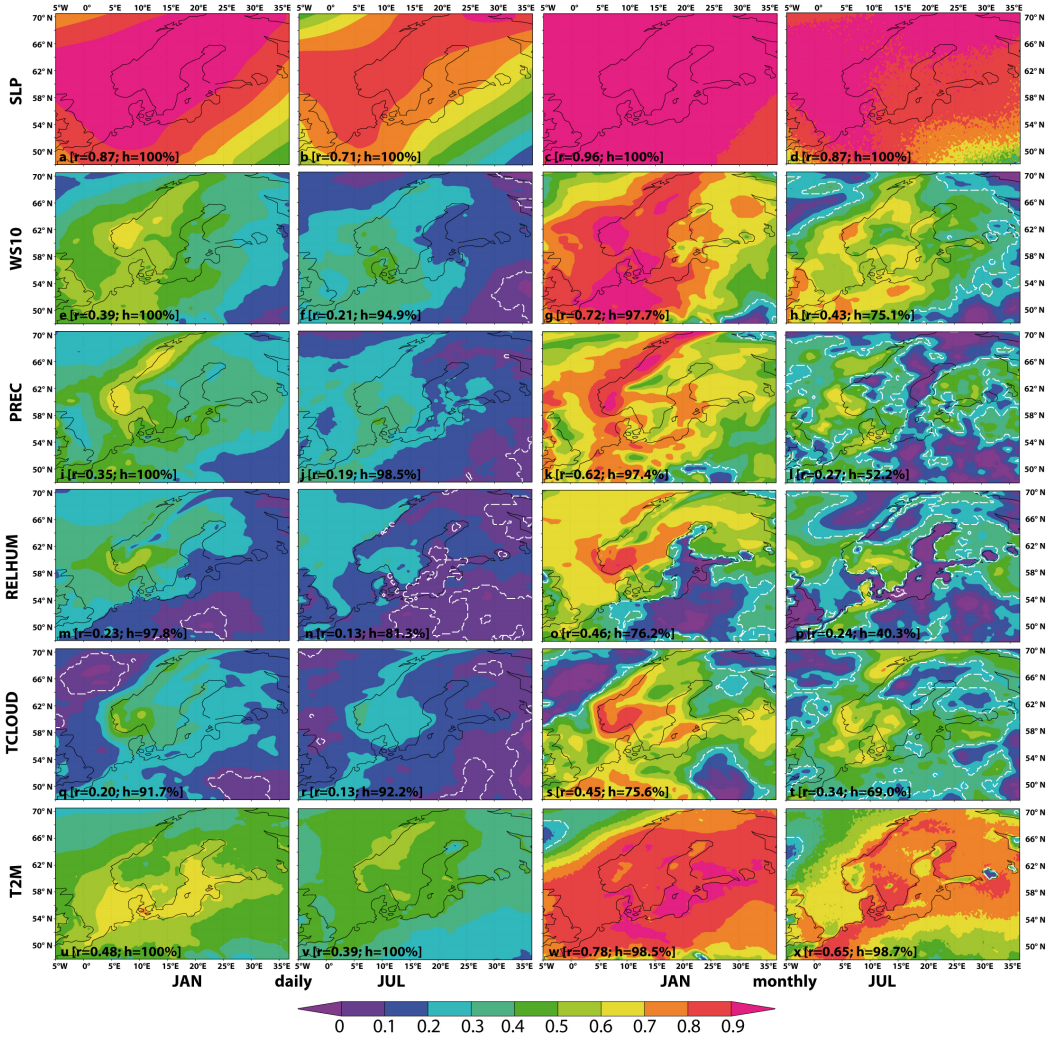

Fig. 6. Correlation maps on daily (left column) and monthly (right column) scale between HiResAFF and the reference fields of RCAO (RCAX for T2M) for January and July 1958–2007. White shaded lines indicate areas where h_0 of zero correlation cannot be rejected with $p < 0.05$.

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 ($\phi \ll 50\,\%$), i.e. 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 ($\phi \ll 50\,\%$), i.e. over the SE of the domain.

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 E-NE show overestimated variance in addition to large parts of the NE-Atlantic in July. On monthly scale, variance is

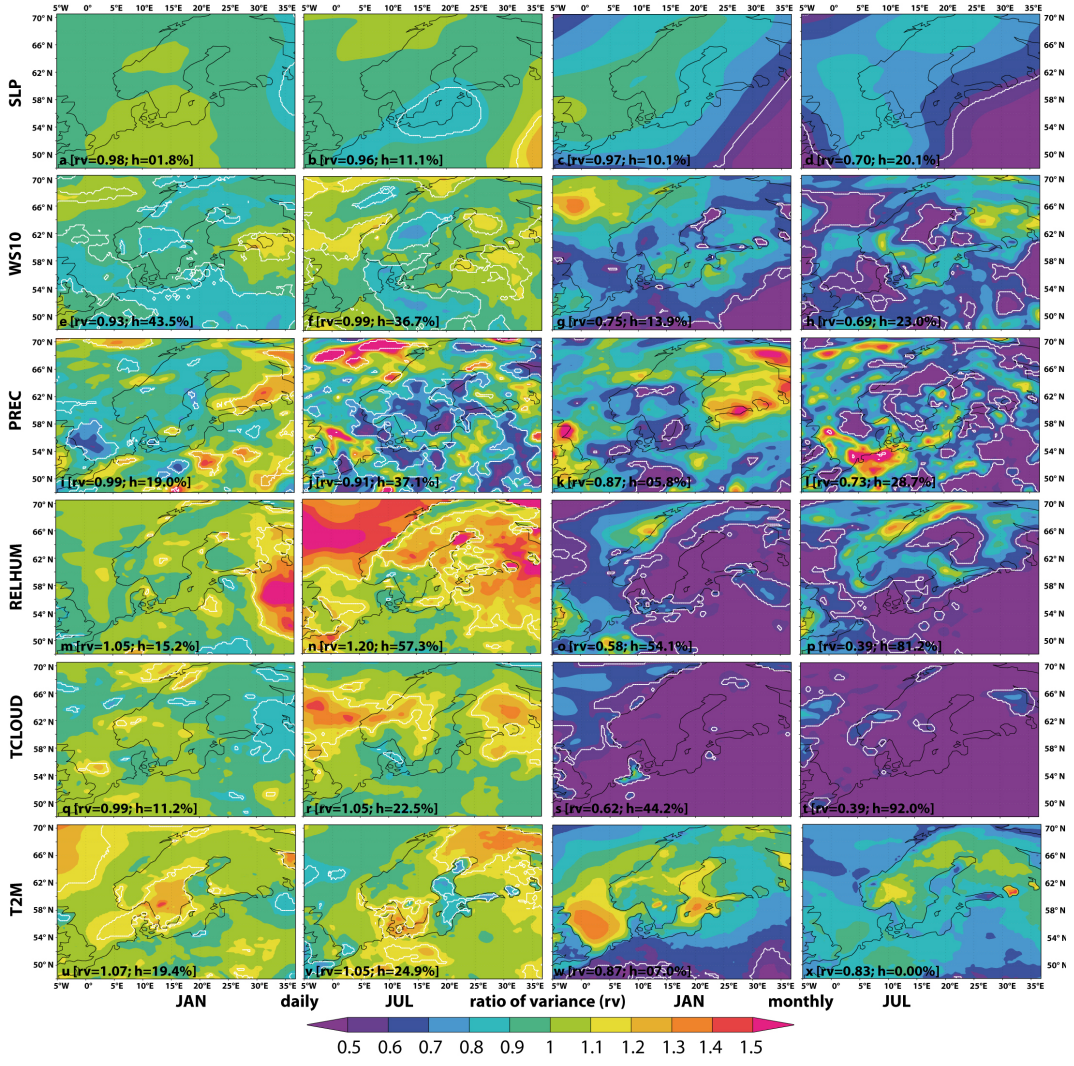

Fig. 7. Ratio of variance on daily (left column) and monthly (right column) scale between HiResAFF and the reference fields of RCAO (RCAX for T2M) for January and July 1958–2007. White shaded lines indicate areas where the reconstruction shows significant deviations in variance with $p < 0.05$.

clearly underestimated, with exception of the northern and SW domain in January. For July, variance is strongly underestimated ($\phi \ll 50\,\%$) 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.

3.3.3 Reconstruction bias

The bias in mean (monthly sum in precipitation case) $\Delta\overline{m} = \overline{m}_{\mathrm{HiResAFF}} - \overline{m}_{\mathrm{RCAO}}$ of the reconstruction is shown in Fig. 8. The average bias of the field $\Delta\overline{m}$ and the number of local tests h [%] showing significant deviations with $p < 0.05$ are summarized in Table 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 (up to +1.4 hPa).

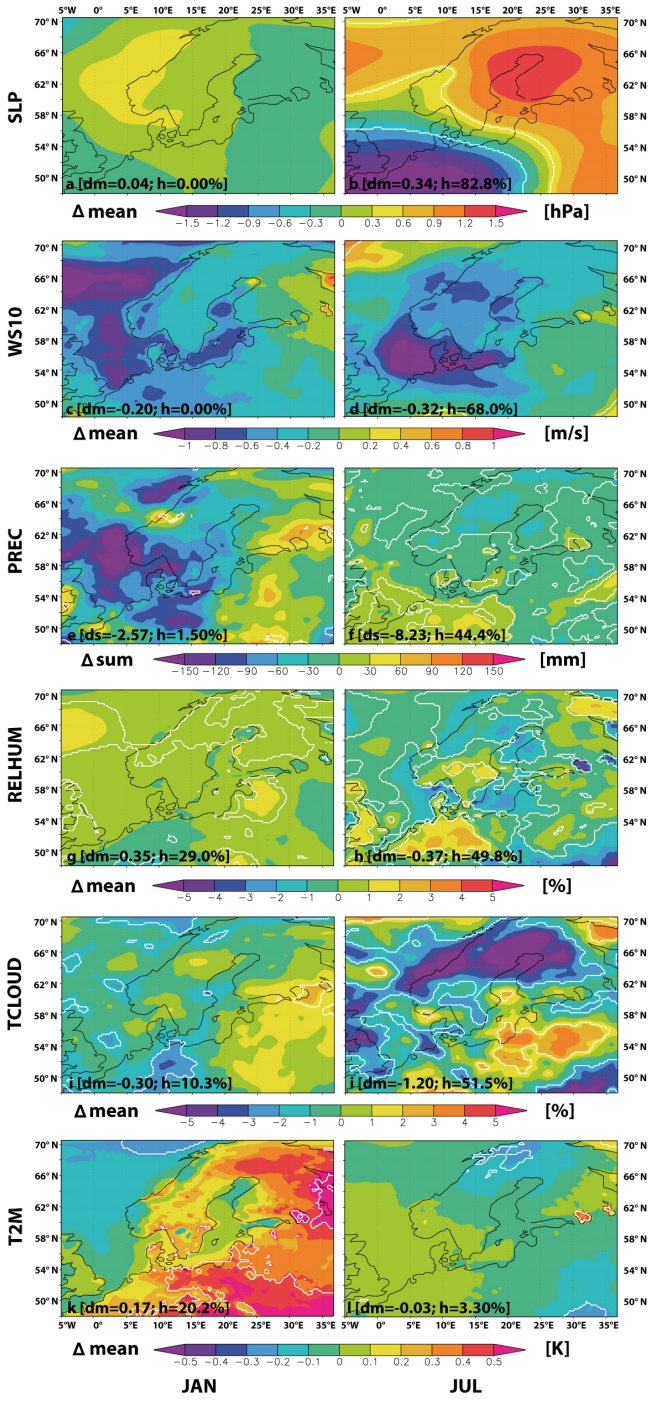

Fig. 8. Mean bias of HiResAFF minus the reference fields of RCAO (RCAX for T2M) for January (left column) and July (right column) 1958–2007. White shaded lines indicate areas where the reconstruction shows significant bias in mean with $p < 0.05$.

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.

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. 2.5) 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. 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.

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. 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,

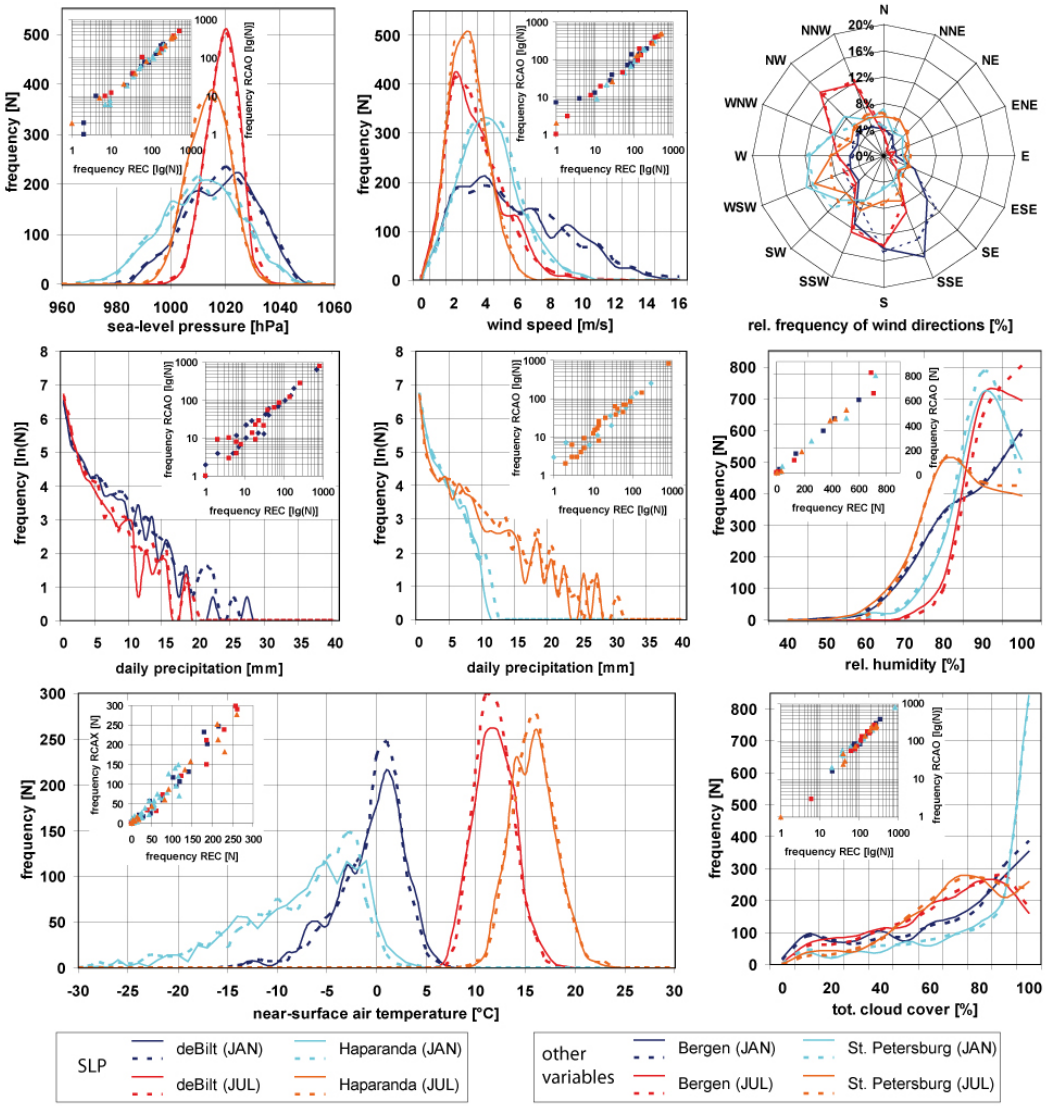

Fig. 9. Comparison of the frequency distributions between reconstruction (HiResAFF, solid lines) and reference fields from RCAO (RCAX for T2M) (shaded lines) for different variables for January and July. Grid points are chosen in the vicinity of de Bilt and Happaranda in thr SLP case to highlight latitudinal 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).

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. 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, 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. 2.3.4) 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. 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 ($T < -5\,^{\circ}\mathrm{C}$) temperatures does not deviate considerably, the frequency of temperatures between $-5\,^{\circ}\mathrm{C}$ and $0\,^{\circ}\mathrm{C}$ 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. 9, largest deviations occur around the mean value leading to a broader distribution of the reconstruction compared to the reference fields.

3.3.5 Auto-correlation

Figure 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.

For daily T2M, two reconstructions are compared in Fig. 10a based on different settings used for the AM. In light blue and orange, the standard-setting T2M reconstruction is shown (Sect. 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. 2.3.4) 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. 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.

4 Discussion

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

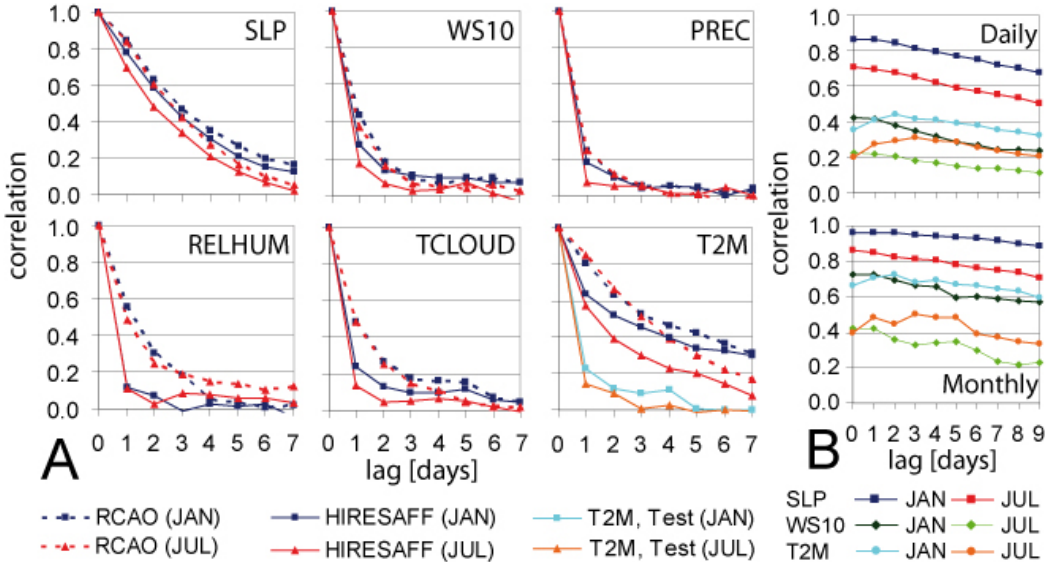

Fig. 10. Effect of implementing persistence in the analog-reconstruction. **(A)** Reconstructed daily auto-correlations of HiResAFF (solid lines) dependent on the used n-lag in the AM with RCAO and RCAX (shaded lines) for January and July at a grid point close to Stockholm. The test cases (light blue, orange) for T2M show the daily serial-correlation if T2M is reconstructed using the standard settings from Sect. 2.3.2. **(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.

a clear loss in the explained variance when linking real data to model fields (Fig. 3). In Sect. 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. 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.

4.2 Robustness of the analog-method

The results in Sect. 3.2.1 and Fig. 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. 3.2.2, Fig. 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. 3.2.3, Fig. 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. 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.

4.3 Validation of HiResAFF

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) 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. 2.3.4).

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.

4.3.2 Variance

Based on the results in Sect. 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. 2.3.4) 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.

4.3.3 Bias

Regarding the deviation in mean (monthly sum for precipitation) of HiResAFF for the different variables (Sect. 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. 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 a 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.

4.3.4 Reconstruction of frequency distributions and autocorrelation

From the results shown in Fig. 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. 2.3.2) 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. 2.3.3) 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. 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.

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. 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. 3 and the discussion in Sect. 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, 2012). 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.

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 50-yr 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. 2.3.4) 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. 2.3.4). 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., 2012) 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.

Acknowledgements. The research leading to these results has 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). We would like to thank Lars Bärring (SMHI), Tuija Ruoho-Airola (FMI), Ari Venäläinen (FMI), Christine Luge (University of Jena) and Gerard van der Schrier (KNMI) for their help to update station data. We also thank Sebastian Wagner (HZG) and Christoph Matulla (ZAMG) for fruitful discussions about the analog-method and two anonymous reviewers for their constructive comments.

Edited by: V. Rath

References

Ansell, T. J., Jones, P. D., Allan, R. J., et al.: Daily mean sea level pressure reconstructions for the European-North Atlantic region for the period 1850–2003, J. Climate, 19, 2717–2742, 2006.

Auer, I., Böhm, R., Jurkovic, A., et al.: 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, 2007.

Barnett, T. and Preisendorfer, R.: 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, 1978.

Bhend, J. and von Storch, H.: 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, 2008.

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

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

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

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

Cheung, K. Y. and Lee, S. M. S: Variance estimation for sample quantiles using the m out of n bootstrap, Ann. Inst. Stat. Math., 57, 279–290, 2005.

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

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

Dee, D. P., Uppala, S. M., Simmons, A. J. et al.: 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, 2011.

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.: Global high resolution versus Limited Area Model climate change projections over Europe: quantifying confidence level from PRUDENCE results, Clim. Dynam., 25, 653–670, doi:10.1007/s00382-005-0052-1, 2005.

Déqué, M., Rowell, D., Lüthi, D., Giorgi, F., Christensen, J., Rockel, B., Jacob, D., Kjellström, E., de Castro, M., and van den Hurk, B.: An intercomparison of regional climate simulations for Europe: assessing uncertainties in model projections, Climatic Change, 81, 53–70, doi:10.1007/s10584-006-9228-x, 2007.

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

Ebisuzaki, W.: 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, 1997.

Efron, B.: The jackknife, the Bootstrap and other resampling plans, J. W. Arrowsmith Ltd., Bristol, England, 1982.

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

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

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

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.: Tropical Pacific – Mid-latitude Teleconnections in Medieval Times, Climatic Change, 83, 241–285, doi:10.1007/s10584-007-9239-2, 2007.

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

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

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

Hurrell, J. W.: 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, 1995.

Jones, P. D. and Moberg, A.: 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, 2003.

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

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.: The NCEP-NCAR 50 year reanalysis, B. Am. Meteorol. Soc., 82, 247–267, 2001.

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., et al.: 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, 2002.

Kruizinga, S. and Murphy, A. H.: 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, 1983.

Livezey, R. E. and Chen, W. Y.: 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, 1983.

Liu, R. Y. and Singh, K.: Moving blocks bootstrap captures weak dependence, in: Exploring the Limits of the Bootstrap, Whiley, 225–248, 1992.

Lorenz, E. N.: 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, 1969.

Matulla, C.: 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, 2005.

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

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

Meier, H. E. M., Eilola, K., and Almroth, E.: 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, 2011a.

Meier, H. E. M., Höglund, A., Döscher, R., Andersson, H., Löption, U., and Kjellström, E.: 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, 2011b.

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., Weslawski, J.-M., and Zorita, E.: 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, 2012.

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

Rockel, B. and Woth, K.: 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, 2007.

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

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

Schimanke, S., Meier, H. E. M., Kjellström, E., Strandberg, G., and Hordoir, R.: 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, 2012

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

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

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

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

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

von Storch, H. and Zwiers, F.: Statistical Analysis in Climate Research, Cambridge Univ. Press, New York, USA, 1999.

von Storch, H., Langenberg, H., and Feser, F.: A spectral nudging technique for dynamical downscaling purposes, Mon. Weather Rev., 128, 3664–3673, doi: 10.1175/1520-0493(2000)128<3664:ASNTFD>2.0.CO;2, 2000.

von Storch, H., Zorita, E., and Cubasch, U.: 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, 1993.

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

Wanner, H., Brönnimann, S., Casty, C., Gyalistras, D., Luterbacher, J., Schmutz, C., Stephenson, D. B., and Xoplaki, E.: North Atlantic Oscillation – Concepts and Studies, Surv. Geophys., 22, 321–382, 2001.

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

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

Zorita, E. and von Storch, H.: 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<2474:TAMAAS>2.0.CO;2, 1999.

Zorita, E., Hughes, J., Lettenmaier, D., and von Storch, H.: 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, 1995.