

Evaluation of a dynamic downscaling of precipitation over the Norwegian mainland

By E. ORSKAUG^{1*}, I. SCHEEL², A. FRIGESSI^{3,1}, P. GUTTORP^{4,1}, J. E. HAUGEN⁵,
O. E. TVEITO⁵ and O. HAUG¹, ¹Norwegian Computing Center, Gaustadallen 23, PO Box 114, Blindern,
0314 Oslo, Norway; ²Department of Mathematics, University of Oslo, PO Box 1053, Blindern, 0316 Oslo, Norway;
³Department of Biostatistics, University of Oslo, PO Box 1122, Blindern, 0317 Oslo, Norway; ⁴Department of
Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA; ⁵The Norwegian Meteorological
Institute, PO Box 43, Blindern, 0313 Oslo, Norway

(Manuscript received 12 November 2010; in final form 3 March 2011)

ABSTRACT

In order to assess the potential of regional climate models to be used to project future weather events, a first step is to study the regional model forced by actual weather, or more precisely by reanalysis of weather data. In this paper we investigate how well the Norwegian regional model HIRHAM, forced by ERA-40 reanalysis data, compares to observed precipitation data from the Norwegian Meteorological Institute over Norwegian mainland. This paper aims to show how standard methods of statistical testing may be used to assess dynamic downscaling. Methods considered are the Kolmogorov–Smirnov two-sample test, a Fisher exact test for equality of quantiles, an Extreme Value Theory test, where equality of the 1-yr return levels are tested, and equality of wet-day frequency. All tests are performed seasonally. The regional model is skillful in describing the lower quartile of the precipitation distribution, but underestimates higher levels of precipitation. Our results indicate that the regional model has too many but too small rain events for all seasons.

1. Introduction

There is a broad scientific consensus to expect changing weather patterns, by type, frequency, intensity, variability and location (for an overview see e.g. Balkema and Haan, 1974; Maraun et al., 2010). The new climate will have a major impact on nature and society, which has to prepare adaptation strategies, in parallel with emission reduction policies. There is urgency in providing information on the impacts of climate change. Climate change simulations are based on complex deterministic models, driven by uncertain forcings that are based either on observations or scenarios. Effects of climate change are downscaled from coarse global models to finer resolution regional models. The purpose of regional climate models is to give stakeholders and decision makers a representation, possibly a reliable projection, at a useful spatio-temporal scale, of future weather events. For example, an insurance company may be interested in precipitation projections under various possible futures to assess the changing risk of damages to buildings or flooding (Scheel and Hinnerichsen, 2010). Typically scenario runs are done with the

regional model forced by a global coupled ocean–atmosphere model. The question then becomes how reliable these regional models are, at the scale needed by the actual effect study. For example, to understand patterns of risk for the insurance of buildings, precipitation at a mesoscale level are needed, say on a $25 \times 25 \text{ km}^2$ grid, at least daily. Uncertainty is an unavoidable aspect of climate modelling. There are uncertainties resulting from insufficient detail, from incomplete data, from numerical instability, from approximations and from measurement errors. The regional model is then thought of as performing a dynamic downscaling of the global model. Since a global model is not an accurate description of weather, but rather of the distribution of weather, it is reasonable, when attempting to assess the accuracy of a regional model, to force it with the actual state of the atmosphere. Of course, we do not know that, but can at least estimate it by a reanalysis of climate data. This is sometimes called hindcasting (Rummukainen, 2010). Here we look at the regional model HIRHAM (Haugen and Iversen, 2008) forced by the ERA-40 reanalysis (Uppala et al., 2005). To understand how well this regional model performs, we compare it to observation data from the Norwegian Meteorological Institute. This comparison can highlight specific deficiencies of regional models for the Norwegian Peninsula, where forecasts of precipitation are notoriously difficult due to the topography of the country. This

*Corresponding author.

e-mail: elisabeth.orskaug@nr.no

DOI: 10.1111/j.1600-0870.2011.00525.x

will pave the way for appropriate non-linear, time and space inhomogeneous corrections, which can subsequently be applied to downscaling of future climate.

Like weather forecasts in general, being just an incomplete representation of the physics involved, the downscaling introduces inaccuracies and errors in the resulting weather variables (Christensen et al., 2010). The scale of the errors varies geographically depending on the current state of the atmosphere. Feeding the model with the best representation of the atmosphere along the boundary of the integration area, emphasizes the errors of the downscaling process itself and minimizes the contribution added from propagating discrepancies inherently present in the boundary ERA-40 data. Reliant on the downscaling, the $25 \times 25 \text{ km}^2$ downscaled ERA-40 data are still supposed to possess properties similar to real weather locally over longer time periods.

This paper offers a fresh comparison of the distributions of downscaled ERA-40 precipitation data with triangulated precipitation observations, over 40 yr of weather over Norwegian mainland. In Section 2 we describe the data precisely. In Section 3 we present the methodology used to compare two distributions globally and in specific parts of their range, for example, the wet-day frequency, intermediate quantiles and returns representing extreme precipitation, using both parametric and non-parametric tools. Results are presented in Section 4. The final section presents our conclusions. We point the reader to our supplementary material (Orskaug et al., 2010) for additional comparisons.

2. Data

The data used in this study constitute 40 yr of daily precipitation values for the Norwegian mainlands, covering the period 1 January 1961 to 31 December 2000. The data set is twofold, where one part consists of downscaled ERA-40 reanalysis model data and the other is based on observations.

2.1. ERA-40 reanalysis data

Broadly speaking, reanalysis data express the best estimate available for the current state of the atmosphere. They are formed in retrospect from feeding various sources of meteorological observations into a computerized atmospheric model that smoothes the observations and brings them into consistency. Consistency is achieved by re-running with a frozen version of the atmospheric model, although the data types and coverage have changed considerably during the ERA-40 period, in particular the introduction of satellite measurements.

ERA-40 reanalysis data are a product of the ECMWF (European Center for Medium-Range Weather Forecasts) in the United Kingdom. Measurements obtained from balloons, aircraft, buoys, radiosondes, scatterometers and satellites are propagated through ECMWF's computer model. The model out-

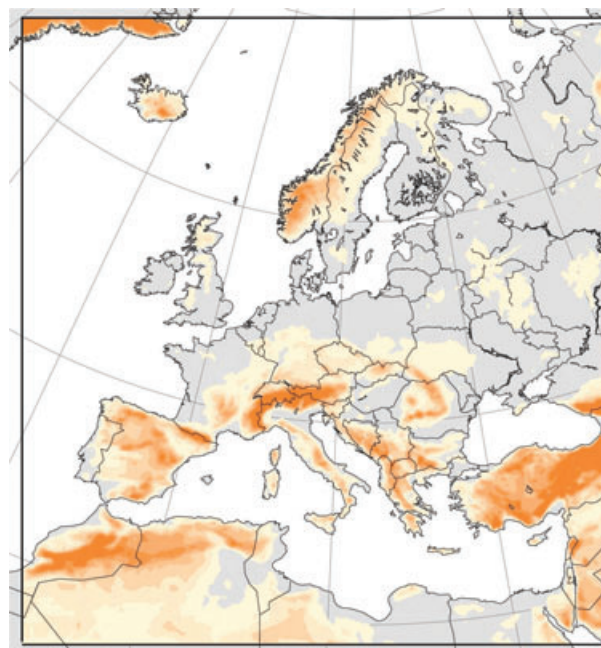


Fig. 1. Integration area for the ERA-40 data.

puts the state of the atmosphere through daily meteorological variables on a grid. Our data were extracted on a regular latitude–longitude grid with a horizontal spacing close to the actual resolution of the atmospheric model ($125 \times 125 \text{ km}^2$).

Downscaled ERA-40 reanalysis data are collected from the ENSEMBLES project website (Christensen et al., 2010). Gridded ERA-40 data along the boundary of an integration area covering most of Europe (Fig. 1) are dynamically downscaled to weather variables on a grid with a spatial resolution of $25 \times 25 \text{ km}^2$, which amounts to 777 grid cells covering Norwegian mainland. The downscaled ERA-40 reanalysis data (Haugen and Haakenstad, 2006) will be referred to as dERA40 in this paper. The downscaling is done by the Norwegian Meteorological Institute by means of their HIRHAM Regional Climate Model (Haugen and Iversen, 2008).

2.2. Observation based precipitation data

Precipitation is observed by stations irregularly distributed across Norway. Based on all observations of precipitation available at every time, high-resolution precipitation grids are estimated applying a spatial interpolation method. Interpolation of precipitation is a challenging task. Precipitation is a non-continuous process in time and space and is therefore not easy to fit to a Gaussian distribution that is a fundamental assumption in spatial statistics. Beside the complex nature of the precipitation process itself, a complex topography, like the Norwegian, makes spatial distribution of precipitation even more challenging. Observations from sites located only a few tens of kilometres

apart may exhibit differences of a factor of 10 in mean annual precipitation.

The method applied to estimate gridded precipitation data at the Norwegian Meteorological Institute is Delaunay triangulation (Jansson et al., 2007). In the procedure applied for Norwegian precipitation, two interpolations for each event are carried out. The observed precipitation is triangulated and a $1 \times 1 \text{ km}^2$ grid is derived from the triangulated surface. A similar grid based on the elevations of the available observation stations is also interpolated. The difference between this elevation grid and a real terrain model is used to adjust the interpolation precipitation for terrain effects using a pre-defined elevation gradient. The terrain adjustment is activated when interpolated precipitation exceeds 0.05 mm.

Precipitation observations are known to be systematically underestimated due to wind loss. The undercatch is particularly large for snow precipitation, where the loss is estimated to be 80% in very wind exposed areas. This phenomenon is partly accounted for by adjusting the observed precipitation by a correction factor (Førland et al., 1996) that depends on the exposure of each measurement location.

To compare with dERA40, the $1 \times 1 \text{ km}^2$ gridded measurements are aggregated into the larger $25 \times 25 \text{ km}^2$ grid. This is obtained by collecting all $1 \times 1 \text{ km}^2$ grid cells with centre points within an ERA-40 cell, and taking their mean as a representation of the measured precipitation inside that grid cell. We use the abbreviation OBS for these data.

3. Methods for comparison

Several different measures are employed to compare downscaled ERA-40 reanalysis model data (dERA40) and observed precipitation. It is well known that model data have problems with spurious precipitation, known as the drizzle effect. A simple approach to avoid the drizzle effect is to condition on a small, positive threshold (Hay and Clark, 2003). For both data sets, a threshold of 0.5 mm d^{-1} is used for the definition of a wet day, as is also suggested by (Haylock et al., 2008) in the ENSEMBLES project.

Most of the measures are conditioned on wet days, that is, dry days are left out. Exceptions are ‘Wet-day frequency’, where still the threshold of 0.5 mm d^{-1} is used to define a wet or dry day, but here dry days are not left out.

The day-to-day correlation is partly lost in dERA40 when downscaling the ERA-40 reanalysis data, hence the distributions have to be compared instead of testing day by day. The measures considered are the Kolmogorov–Smirnov test, tests of equality of quantiles, permutation test of the frequency of wet days and test of extreme events (right tail in the distribution) using a generalized Pareto distribution (GPD). Other local measures such as the equality of the means, standard deviation, maximum 5-d precipitation and maximum number of consecutive dry days are

also considered, and are reported in the supplemental material (Orskaug et al., 2010).

All tests are done seasonally and based on the 777 grid cells that cover Norwegian mainland. We then have $4 \times 777 (=3108)$ tests for each measure. The seasons are winter from December to February, spring from March to May, summer from June to August and autumn from September to November. Under independence between the tests, if all null hypotheses are true, we would expect about 155 spurious significances at 95% confidence level.

3.1. Kolmogorov–Smirnov test

We first compare the two empirical distributions of OBS versus dERA40 for each season and grid cell. To test the null hypothesis that these two distributions are equal, we use the Kolmogorov–Smirnov two-sample test (e.g. Conover, 1971). The Kolmogorov–Smirnov test is a non-parametric test for assessing the equality of the underlying distributions of two data sets. The Kolmogorov–Smirnov statistic quantifies a supremum distance between the empirical distribution functions.

$$D_{n,n'} = \sup_x |F_n(x) - G_{n'}(x)|, \quad (1)$$

where $F_n(\cdot)$ and $G_{n'}(\cdot)$ are the empirical distribution functions of OBS and dERA40, and $\sup_x$ denotes the supremum of the set of distances between these empirical distributions. The null hypothesis, H_0 , is rejected at level α if

$$K_{n,n'} = \sqrt{\frac{nn'}{n+n'}} D_{n,n'} > K_\alpha, \quad (2)$$

where K_α is found from

$$P(K_{n,n'} \leq K_\alpha | H_0) = 1 - \alpha. \quad (3)$$

$K_{n,n'}$ is approximately Kolmogorov distributed when n and n' are large.

To avoid the problem of tied¹ data, a small, random, normally distributed number $[N(0, \sigma^2)]$, where $\sigma = 0.0000001$, is added to each data point.

Next we consider several local measures, that is, measures that consider a specific characteristic of the distribution.

3.2. Quantiles

The equality of quantiles is tested using Fisher’s exact test (e.g. Bickel and Doksum, 1977). Each empirical quantile is calculated from the compound data set including dERA40 and the observations, say $\hat{Q}_x^{\text{tot}}$, where x is a percentage (0.05 and 0.95).² A 2×2 contingency table is made by counting the number of days with precipitation above $\hat{Q}_x^{\text{tot}}$ and below $\hat{Q}_x^{\text{tot}}$, for both

¹ Inherently present in data represented with a finite number of digits.

² Other quantiles are also considered and reported in the supplemental material (Orskaug et al., 2010): $\hat{Q}_{0.10}^{\text{tot}}$, $\hat{Q}_{0.25}^{\text{tot}}$, $\hat{Q}_{0.50}^{\text{tot}}$, $\hat{Q}_{0.75}^{\text{tot}}$, $\hat{Q}_{0.90}^{\text{tot}}$.

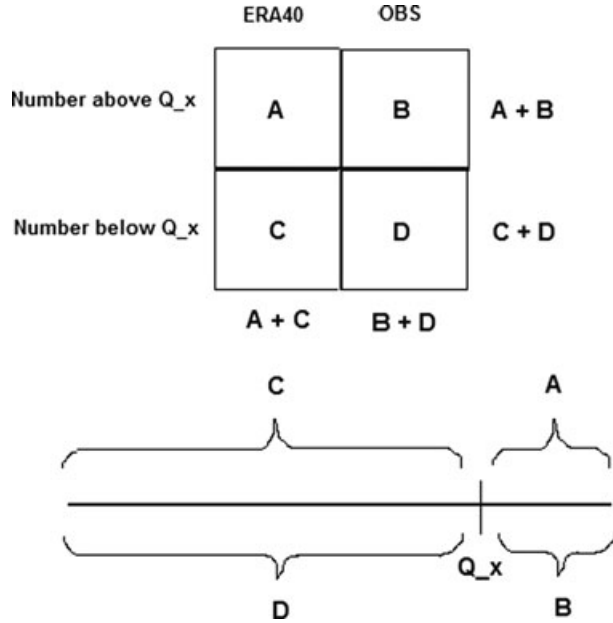


Fig. 2. Construction of the 2×2 contingency table for Fisher's exact test.

dERA40 and the observations. Figure 2 shows a picture of how the 2×2 contingency table is constructed.

Fisher's exact test is based on the fact that the counts in the table have a joint hypergeometric distribution when conditioning on all four marginals in the 2×2 table (i.e. conditioning on the sums $A + C$, $B + D$, $A + B$ and $C + D$). It is natural to condition on both sets of marginals in this case because they are implied by the classification rule for the table. dERA40 has $A + C$ observations, the observation data set has $B + D$ observations, $A + B$ observations are below Q_x^{tot} and $C + D$ observations are above Q_x^{tot} .

3.3. GPD

In certain applications, for example, in the insurance industry, it is imperative to consider tail properties of the data sets, to determine if one is more heavy tailed than the other.

The Pickands-Balkema-de Haan Theorem (Balkema and Haan, 1974) suggests that for sufficiently high thresholds, u , the data exceeding that threshold will follow a Generalized Pareto Distribution (GPD).

The form of the GPD is

$$F_{\xi, \sigma}(y) = P(X - u > y | X > u) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0, \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & \xi = 0, \end{cases} \quad (4)$$

where X is the precipitation, u is a threshold that must be specified, $y = x - u \geq 0$ when $\xi \geq 0$ and $0 \leq y = x - u \leq -\sigma/\xi$ when $\xi < 0$. The two parameters, ξ and σ in (4) are called the shape and scale parameter, respectively.

The parameter u is chosen as the 95%-quantile of the compound dERA40 and observations data set of every single grid cell separately.

To compare the tails of the two data sets we fit separate GPDs to data beyond u , and then use their T -yr return level to assess discrepancies in their right tails. The return level is the GPD-quantile that on average is exceeded once, over T years, where there are 365.25 observations on average per year.

The return level, x_T , is a function of the threshold u and the parameters ξ and σ of the GPD (Brabson and Palutikof, 2000).

$$x_T = u - \frac{\sigma}{\xi} \left(1 - (\lambda T)^{\xi}\right), \quad (5)$$

where T is the return time (in years) and λ is the expected number of peaks above the threshold, u , per year.

T is set to 1 yr and λ is calculated from the data using the 40-yr mean number of peaks above u as an estimate. The quantile x_T is calculated for both data sets.

An intercomparison of the return levels x_T is accomplished through permutation testing (e.g. Good, 2005). Let x_T^{dERA40} and x_T^{OBS} denote the T -yr return levels of the dERA40 and OBS data sets, respectively, and introduce $\Delta_T = x_T^{\text{dERA40}} - x_T^{\text{OBS}}$. Now, testing $H_0: x_T^{\text{dERA40}} = x_T^{\text{OBS}}$ against $H_1: x_T^{\text{dERA40}} \neq x_T^{\text{OBS}}$ is equivalent to testing $H_0: \Delta_T = 0$ against $H_1: \Delta_T \neq 0$.

The permutation test is carried out for each grid cell separately, and compares the observed test statistic $\Delta_{T, \text{OBS}}$ against a sample set of size $B = 1000$ from the approximate null distribution of Δ_T . Reference sample data sets are formed from swapping, with 50% probability, precipitation values of the original dERA40 and OBS data sets on a daily basis, leaving complementary return levels $\{x_{T,b}^{\text{dERA40}}\}_{b=1}^B$ and $\{x_{T,b}^{\text{OBS}}\}_{b=1}^B$. Corresponding differences $\{\Delta_{T,b}^{\text{dERA40}}\}_{b=1}^B$ are formed from $\Delta_{T,b} = x_{T,b}^{\text{dERA40}} - x_{T,b}^{\text{OBS}}$. The p -value of the test is calculated from the proportion of $\Delta_{T,b}$'s, $b = 1, \dots, B$, in absolute value greater than $\Delta_{T, \text{OBS}}$.

3.4. Wet-day frequency

The wet-day frequency is the proportion of wet days (among all days in the data). The equality of the wet-day frequencies is tested by permutation testing. This is done by resampling the two data sets 1000 times, exchanging daily dERA40 and observation figures with a 50% probability. The null hypothesis is that dERA40 and the observations have the same wet-day frequency, and the p -value is calculated in a similar way as described in Section 3.3.

4. Results

Figures 4–7 show p -values of the different tests as spatial maps over Norway. When there are no significant differences, a grid cell is yellow. For p -values less than or equal to the level of significance, α , grid cells are either red or blue, dependent on the direction of the difference. The darker the colour of a grid



Fig. 3. Kolmogorov–Smirnov two-sample test used with a significance level $\alpha = 0.05$. When there are no significant differences, a grid cell is yellow, that is, $p\text{-value} > \alpha$. If the $p\text{-value} \leq \alpha$ the grid cell is blue.

cell, the greater the difference in absolute value, as indicated in the legends.

Based on the Kolmogorov–Smirnov test we reject the null hypothesis of equality of the distribution of OBS and dERA40 in almost all the grid cells for all the four seasons, see Fig. 3. There are some yellow grid cells indicating equality in the distributions, but these seem more or less randomly scattered. Typically, in situations with lots of data, even small differences come out statistically significant. Hence the Kolmogorov–Smirnov test

easily returns rejection of the null hypothesis. To illustrate how different the distributions of dERA40 and OBS are with other tools than Kolmogorov–Smirnov, empirical density functions and empirical cumulative distribution functions of four grid cells are shown in the supplemental material (Orskaug et al., 2010). The empirical density functions and empirical cumulative functions visually agree with the Kolmogorov–Smirnov test. The global picture for all the seasons is that dERA40 does not have the same distribution as the observations. A detailed

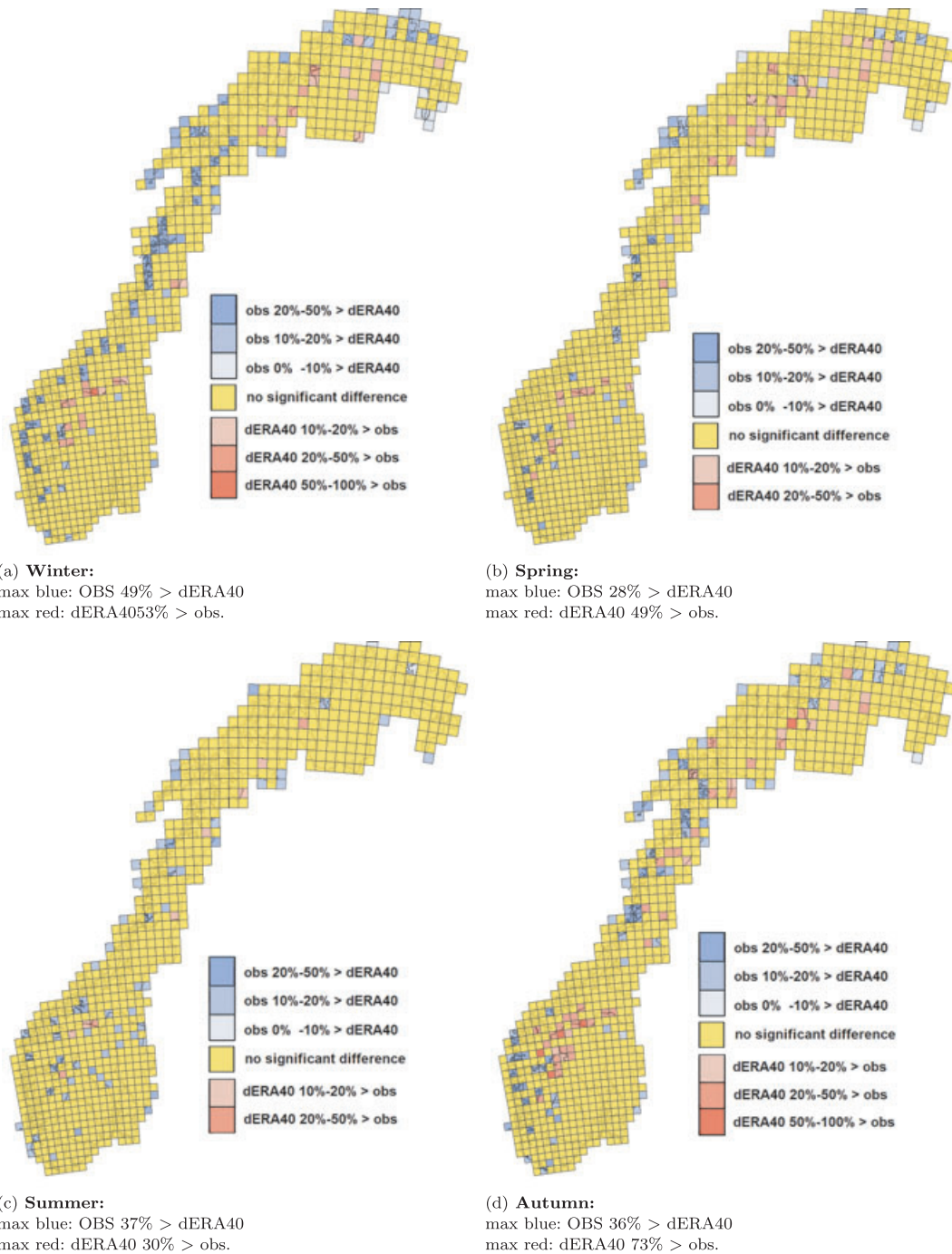


Fig. 4. Test of equality in the 0.05 quantile (conditioned on wet days). The plots show results for significance level $\alpha = 5\%$. Blue colour indicates that $Q_{0.05}^{\text{dERA40}} < Q_{0.05}^{\text{OBS}}$, significantly, and red colour indicates that $Q_{0.05}^{\text{dERA40}} > Q_{0.05}^{\text{OBS}}$, significantly.

comparison of specific local features of the two distributions may reveal significant differences in certain parts.

The results for the smaller quantile (0.05) displayed in Fig. 4, are quite different from the high quantile (0.95) shown in Fig. 5. For the 0.05 quantile there are few rejections, hence dERA40 appears to reproduce the distribution of the observa-

tions well season- and nationwide for low precipitation (but still $> 0.5 \text{ mm d}^{-1}$). More quantiles have been considered in the supplementary material (Orskaug et al., 2010), and it seems that dERA40 reproduces the distribution of the observations well below the first quantile. For the 0.95 quantile, however, the overall picture is mainly blue grid cells, although non-rejections are

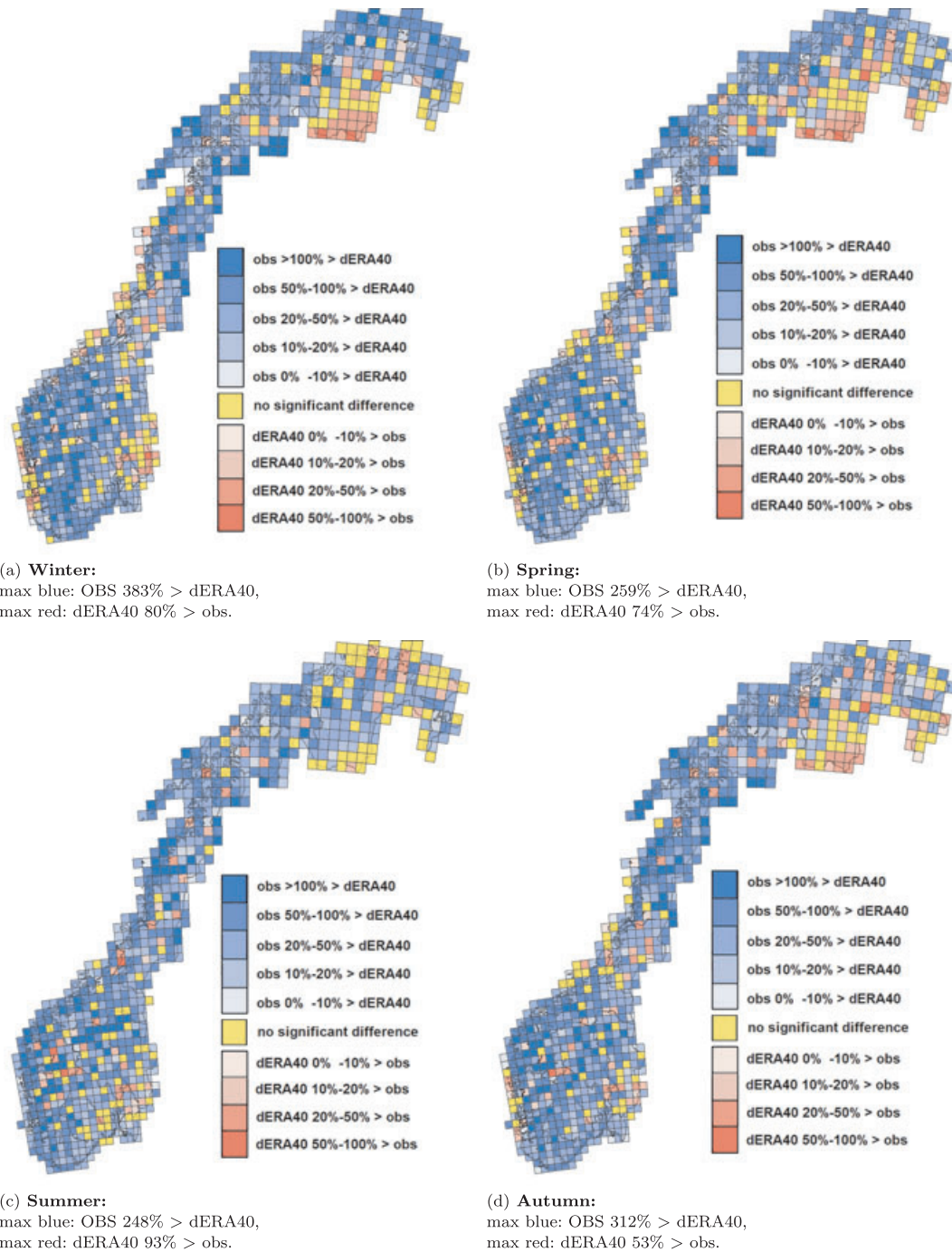


Fig. 5. Test of equality in the 0.95 quantile (conditioned on wet days). The plots show results for significance level $\alpha = 5\%$. Blue colour indicates that $Q_{0.95}^{\text{dERA40}} < Q_{0.95}^{\text{OBS}}$, significantly, and red colour indicates that $Q_{0.95}^{\text{dERA40}} > Q_{0.95}^{\text{OBS}}$, significantly.

found in the county Finnmark, along the outermost belt of the west coast (except in summer) and in areas around the southeastern border to Sweden. This tells us that dERA40 underestimates high precipitation. In Finnmark and in inner parts of southern Norway there is a cluster of some red grid cells (although not that pronounced in summer), indicating that the distributions differ

with the dERA40 quantile being higher than that of the observations in this area. The different behaviour of the error in the inner parts of Finnmark can be explained from a dominating very cold and relatively dry climate compared with the milder and wetter coastal surrounding areas, where the regional model is underestimating the higher tail for wet areas and overestimating for dry

and very cold areas. A similar argument may be used to explain the behaviour for the inner parts of south east during winter. For coastal areas in west of Norway, the differences are influenced by orographic effects, where lifting of air masses from west and release of precipitation are not captured with the same level of details in dERA40 in an area dominated by very steep terrain.

Summing up the quantile test results, there is evidence that the left part of the distribution of dERA40 behaves reasonably well, whereas substantial discrepancies remain in the right tail.

Even though the Kolmogorov–Smirnov test indicates differences in the distributions, local features in the distributions might still be equal. Figure 6 illustrates the comparison between return

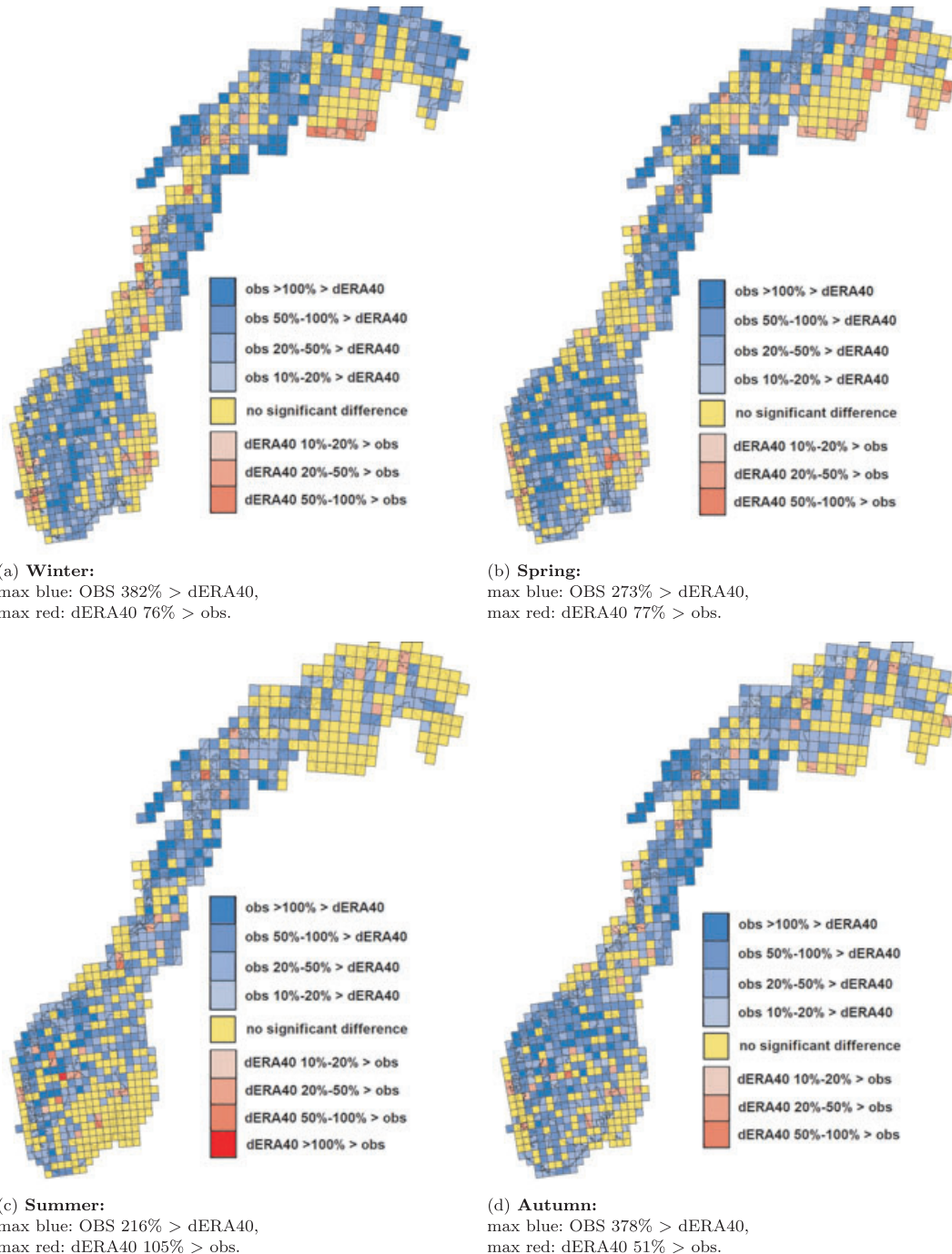


Fig. 6. 1-yr return level tests based on fitting GPDs to data above the 0.95 quantile of the compound data set of dERA40 and observations. The plots show results for significance level $\alpha = 5\%$. Blue colour indicates that $x_T^{\text{dERA40}} < x_T^{\text{OBS}}$, significantly, and red colour indicates that $x_T^{\text{dERA40}} > x_T^{\text{OBS}}$, significantly.

levels of dERA40 versus OBS. Return levels corresponding to $T = 1$ yr return periods and $u = 0.95$ quantile of the compound data set of dERA40 and OBS (in a grid cell) are assessed. We have also tested other values of T (to 0.5 and 2) and u (to 0.90 quantile and 0.975 quantile), but the picture does not change perceptibly for any of these options. Results indicate that 1-yr return levels are more similar than expressed through the Kolmogorov–Smirnov test, that is, although no indication of a perfect match, the right-end part of the distributions are more similar than the full distributions. For the tails of a distribution, where data are more scarce, parameter estimates are more uncertain and thus, more confidence is needed to reject the null hypothesis of equality of return levels. Yellow grid cells are predominant in Finnmark in all seasons, though less in autumn, as well as along the western coast and penetrating up to about 100 km in mid-Norway. Also, we see fair return level predictions in southeastern Norway including Oslo, particularly in summer.

Especially in the inner mountainous parts of southern Norway, non-coastal regions of the county Nordland and along the coast from Lofoten (at the west coast in the county Troms) and northwards, there are rejections of the hypothesis that the right tails of the two distributions are the same. This is indicated by mostly blue grid cells in these areas, meaning that the return level of dERA40 is less than that of the observations. Observe that the return levels can be underestimated by up to 50–100% by dERA40. This confirms the findings of the 0.95 quantile maps of Fig. 5, namely that dERA40 has a tendency to underestimate large daily precipitations, but return level predictions, especially useful for adaptation and mitigation strategies, are not too far off over substantial areas of Norway.

Finally, we compare the frequencies of wet days predicted by dERA40 and OBS. Figure 7 shows that there are mostly rejections for all seasons. The total picture for all the seasons is that the wet-day frequency of dERA40 is greater than for the observations. This indication is most notable in spring for the eastern part of Norway and in Finnmark. An exception to this is the tendency to more rainy days in the observations than those captured by dERA40 during summer in the southeastern parts of the country than the rest of the year. Wet-day frequencies have also been assessed for thresholds of 0.1, 1 and 2 mm precipitation. When the threshold is set to 0.1, the pattern of Fig. 7 is more or less retained, but the intensity of the colours decreases. This might be that for a decreasing threshold, the number of wet days increases for both dERA40 and OBS, leaving a reduced relative ratio of dERA40 to OBS. For a threshold equal to 1, there are only minor differences from Fig. 7, expressed through a few more yellow and blue grid cells. This indication is more distinct for a threshold equal to 2, where the number of red grid cells are further reduced, and yellow and blue grid cells become equally influential (especially in summer). This could be due to a majority of the drizzle effect days of dERA40 now being classified as dry days instead of wet days. The higher wet-day frequency of the dERA40 data does not contradict the outcome

of the quantile tests. Rather, dERA40 tends to scatter its total rainfall amount temporally without reaching the high levels of precipitation registered for certain days in the observations.

A transformation of dERA40 could be thought of as a solution to improve the reproduction of the observations. Simple transformation, using a linear regression, was tried out, but this did not improve the results [reported in the supplemental material (Orskaug et al., 2010)]. A more complex transformation is needed, and we are currently experimenting with non-linear, inhomogeneous corrections in space.

5. Conclusions

The regionally downscaled ERA-40 model data show good agreement with the observations when the rainfall amounts are small (below the first quartile). For rainfall amounts beyond this, the agreement disappears. Our results indicate that the regional model has too many and too small rain events. This is consistent with ERA40 showing a negative precipitation bias over much of Europe, attributed to too little precipitation in connection with extratropical cyclonic systems (Hagemann et al., 2010). ERA40 is also, due to its large grid size, unable to correctly describe the orographic precipitation near the Norwegian coastline. These features are also consistent with the analysis of 20-yr return values of precipitation in the Swedish regional model RCA3 compared to the high-resolution reanalysis E-OBS (Nikulin et al., 2011), which found a high bias in Norway in the summer, and a substantial low bias in the winter. The former is most likely due to the difficulty of simulating convective precipitation.

The major aim of this paper is to show how standard statistical testing may be used to assess dynamical downscaling, and thereby contribute to the purpose of standardizing an intercomparison tool for validating downscaling methods. However, the evaluation could be extended to other models and data sets, for example, from the ENSEMBLES project, to produce a comprehensive set of performance indicators. One alternative would be to compare various GCMs, with a common downscaling technique. Comparing each such downscaled model with the ground truth data tells which GCM would come closest to the observed precipitation patterns. In this way we might be able to identify different shortcomings, specific for each GCM. In principle, one could then patch together these downscaled predictions, to create a synthetic but more accurate prediction. Another possibility is to compare directly downscaled data, each from a specific GCM. A third possible avenue for future research is to compare various downscaling approaches applied to the same GCM, in order to establish differences between these.

For the purpose of using regional models to project precipitation in future scenarios, this analysis shows that there is still a long way to go. The current model is unable to produce realistic rainfall at the detailed level needed for regional planning and other impact studies. The next step in this research will be to

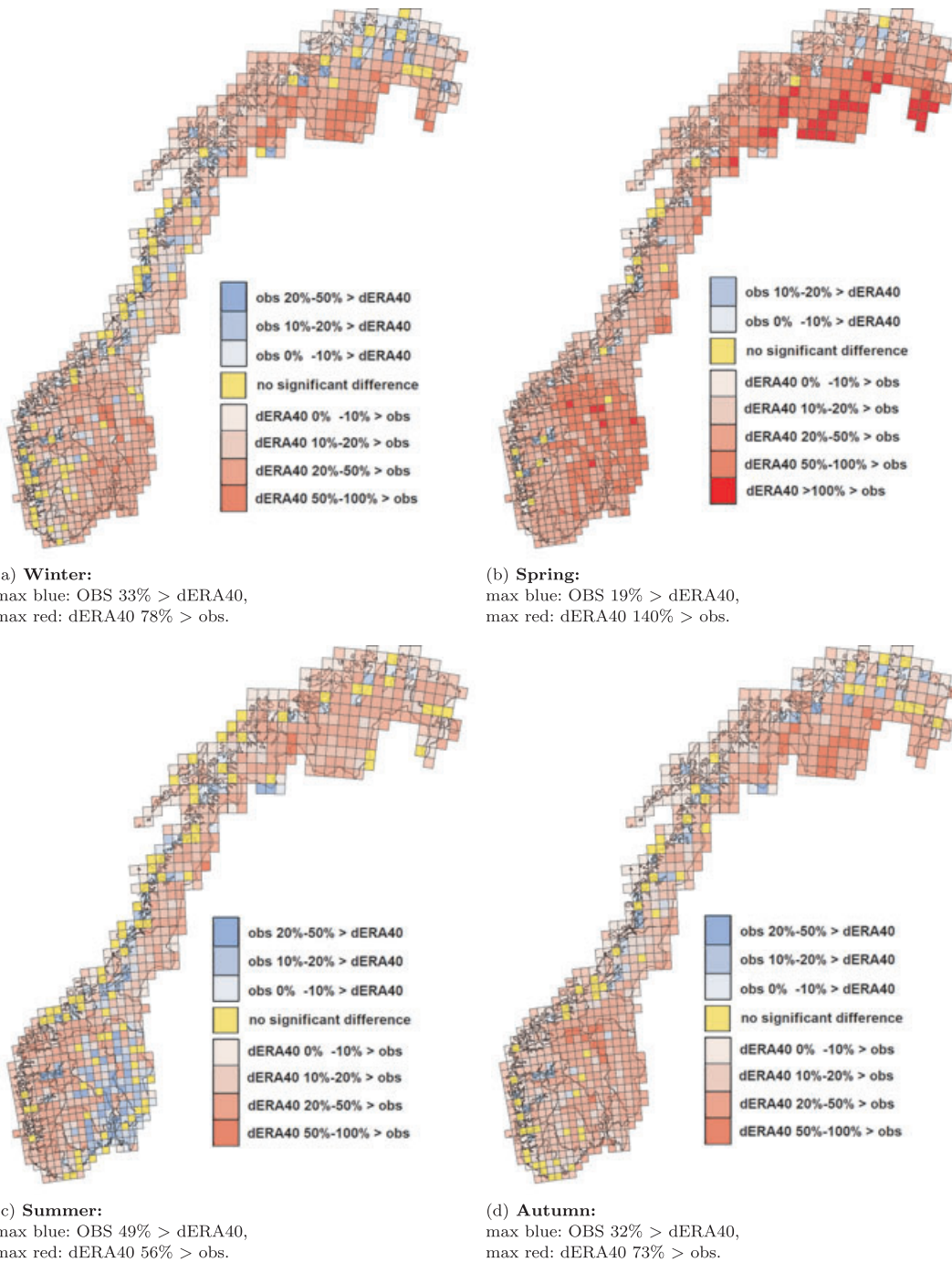


Fig. 7. Test of equality in the wet-day frequency. The plots show results for significance level $\alpha = 5\%$. Blue colour indicates that $\text{freqWet}^{\text{dERA40}} < \text{freqWet}^{\text{OBS}}$, significantly, and red colour indicates that $\text{freqWet}^{\text{dERA40}} > \text{freqWet}^{\text{OBS}}$, significantly.

develop a spatially indexed set of transformations that will make the hindcasting distribution closer to the observed data.

6. Acknowledgments

Funding for writing of this manuscript is provided through the Insuring Future Climate Change project granted by the Research Council of Norway (contract no. 193708/S30), and Statistics for

Innovation (*sfi*)². We thank Henry Wynn and Lenny Smith at the Centre for the Analysis of Time Series, London School of Economics and Political Science, for helpful discussions.

References

- Balkema, A. and Haan, L. 1974. Residual life time at great age. *Ann. Prob.* **2**, 792–804.

- Bickel, P. J. and Doksum, K. A. 1977. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco.
- Brabson, B. B. and Palutikof, J. P. 2000. Tests of the generalized Pareto distribution for predicting extreme wind speeds. *J. Appl. Meteor.* **39**, 1627–1640.
- Christensen, J. H., Kjellström, E., Giorgi, F., Lenderink, G. and Rummukainen, M. 2010. Assigning relative weights to regional climate models: exploring the concept. *Clim. Res.* **44**, 179–194. doi:10.3354/cr00916.
- Conover, W. J. 1971. *Practical Nonparametric Statistics*. John Wiley & Sons, New York.
- Førland, E. J., Allerup, P., Dahlstrøm, B., Elomaa, E., Jonsson, T. and co-authors. 1996. Manual for operational correction of nordic precipitation data. In: *DNMI Klima Report*, Volume 24/96. The Norwegian Meteorological Institute, Oslo, Norway, 66 pp.
- Good, P. I. 2005. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer, New York.
- Hagemann, S., Arpe, K. and Bengtsson, L. 2010. Validation of the hydrological cycle of ERA-40. In: *ECMWF ERA-40 Project Report Series*, Volume 24. European Center for Medium-range Weather Forecasts, Shinfield, Reading, UK, 42 pp.
- Haugen, J. E. and Haakenstad, H. 2006. The development of HIRHAM version 2 with 50km and 25km resolution. RegClim General Technical Report, Volume 9. The Norwegian Meteorological Institute, Oslo, Norway, pp. 159–173.
- Haugen, J. E. and Iversen, T. 2008. Response in extremes of daily precipitation and wind from a downscaled multi-model ensemble of anthropogenic global climate change scenarios. *Tellus* **60A**, 411–426. doi:10.1111/j.1600-0870.2008.00315.x.
- Hay, L. E. and Clark, M. P. 2003. Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. *J. Hydrol.* **282**, 56–75.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D. and co-authors. 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.* **113**. doi:10.1029/2008JD010201.
- Jansson, A., Tveito, O. E., Pirinen, P. and Scharling, M. 2007. NORD-GRID: a preliminary investigation on the potential for creation of a joint Nordic gridded climate dataset. Climate, Met.No Report, Volume 3. 48 pp.
- Maraun, D., Wetterhall, F., Ireson, M., Chandler, R. E., Kendon, E. J. and co-authors. 2010. Precipitation downscaling under climate change. Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48**, doi:10.1029/2009RG000314.
- Nikulin, G., Kjellström, E., Hansson, U. and Ullerstig, A. 2011. Evaluation and future projections of temperature, precipitation and wind extremes over Europe in an ensemble of regional climate simulations. *Tellus*, **63A** 41–55. doi:10.1111/j.1600-0870.2010.00466.x.
- Orskaug, E., Scheel, I., Frigessi, A., Guttorp, P., Haugen, J. E. and co-authors. 2010. Supplemental material to: evaluation of a dynamic downscaling of precipitation over Norway. Report SAMBA/50/10. Norwegian Computing Center, Oslo, Norway.
- Rummukainen, M. 2010. State-of-the-art with regional climate models. *WIREs Clim. Change* **1**, 82–96.
- Scheel, I. and Hinnerichsen, M. 2010. The impact of climate change on insurance risk: a study of the effect of climate change scenarios in Norway. Report SAMBA/52/10. Norwegian Computing Center, Oslo, Norway.
- Uppala, S. M., Kållberg, P. W., Simmons, A. J., Andrae, U., da Costa Bechtold, V. and co-authors. 2005. The ERA-40 re-analysis. *Q. J. R. Meteorol. Soc.* **131**, 2961–3012. doi:10.1256/qj.04.176.