

On downscaling probabilities for heavy 24-hour precipitation events at seasonal-to-decadal scales

By RASMUS E. BENESTAD* and ABDELKADER MEZGHANI, *The Norwegian Meteorological Institute, PO Box 43, NO-0313 Oslo, Norway*

(Manuscript received 8 September 2014; in final form 26 February 2015)

ABSTRACT

We examine a method for predicting variations in the probabilities and occurrence of intense local 24-hour precipitation events over seasonal, annual, and 5-yr intervals. The wet-day mean μ can be used to describe the exponential distribution for the wet-day amounts, and provides a basis for forecasting the likelihood of seeing an event above a given threshold. A regression-based downscaling model is calibrated on annual μ and wet-day frequency f_w , respectively, and then tested on seasonal to multi-annual scales. The probabilities for heavy precipitation statistics are taken as the product between a fitted cumulative exponential distribution for the wet-day amounts μ and f_w . The annual number of heavy precipitation events is estimated from the annual probability, where a 90% confidence interval on the number is computed using the 5 and 95 percentiles of a binomial distribution for the predicted probability. The analysis identifies a strong link between large-scale predictors such as mean sea-level pressure or surface temperature and the wet-day frequency. There was a weaker dependency of the wet-day mean to large-scale predictors, which suggests that local-scale processes have a stronger influence on the 24-hour precipitation amounts. The results also suggest that similar physical processes influence the wet-day mean μ on different timescales except for during winter, and that similar physical processes may explain the wet-day frequency on seasonal to annual time scales. The implication is that models calibrated on annual data samples in some cases can give skilful predictions on seasonal time scales.

Keywords: Precipitation, wet-day mean, wet-day frequency, seasonal-to-decadal forecast, validation, statistical downscaling

To access the supplementary material to this article, please see Supplementary files under 'Article Tools'.

1. Introduction

Can we make skilful forecasts of precipitation statistics seasons to decades ahead, given large-scale patterns in surface temperature and mean sea level pressure (SLP)? Local climate information is important for the users in addition to their credibility, salience, and legitimacy. Barsugli et al. (2013) proposed that establishing the credibility of downscaled climate data is more than just evaluation against historical observations. Skilful forecast on lead times from seasons to decades would provide a valuable basis for decision making for many sectors in our society (Palmer et al., 2000). For instance, such forecasts

could guide farmers on choices regarding when and what to sow (Doblas-Reyes et al., 2007), hydroelectric power producers on how much water to retain in the reservoirs, and manufacturers on the volume of production. Furthermore, early warnings on extreme precipitation can help mitigating flooding and provide valuable time for making necessary preparations. One question is whether such early warning systems are possible, even if we could predict the general state and flow of the oceans and the air masses. Here we try to answer this question in more specific terms: whether it is possible to predict local 24-hour precipitation statistics for a season, year, or a decade, given a representative description of the regional large-scale conditions. We also test the hypothesis whether dependencies between large-scale conditions and local rainfall found on annual timescales are valid for seasonal or decadal timescales.

*Corresponding author.
email: rasmus.benestad@met.no

1.1. Background

Several studies in the past have investigated the possibility of skilful seasonal forecasts for large-scale regional conditions based on general circulation models (GCMs) and statistical forecasting models (Barnett and Preisendorfer, 1987; Barnston et al., 1994; Palmer and Anderson, 1994; Dix and Hunt, 1995; Johansson et al., 1998; Stockdale et al., 1998; Derome et al., 2001, 2005; Rodwell and Doblas-Reyes, 2006). Some studies have suggested that a blend of results from different types of GCMs may improve the forecast skill (Krishnamurti et al., 1999; Yun et al., 2003); however, the question whether such blending results in a real enhancement of skill has also been questioned (Weigel et al., 2008).

Seasonal forecasting has been far more challenging for the North Atlantic and European region than for the Tropics, where the El Niño Southern Oscillation (ENSO) is a major factor. There have been some studies on teleconnections between ENSO and Europe (Lloyd-Hughes and Saunders, 2002), but so far with little success in enhancing seasonal forecasting model skill. A major effort was undertaken to develop a European multimodel ensemble system for seasonal to inter-annual prediction, DEMETER (Palmer et al., 2004) and ENSEMBLES (Linden and Mitchell, 2009). A large part of the seasonal variability over Europe can be linked to variations in the North Atlantic Oscillation (NAO), the jet stream, and planetary Rossby waves; hence, skilful seasonal forecasts need to forecast their future state (Derome et al., 2005; Johansson, 2007). Other factors that may have an effect on the subsequent season over Europe may include the snow cover or the Arctic sea ice (Ueda et al., 2003; Benestad et al., 2010; Orsolini et al., 2011). However, it is difficult to find a systematic effect for some of these conditions. Furthermore, their effect may vary with season. It is also possible that the stratosphere may play a role for seasonal variations (Baldwin and Dunkerton, 2001; Thompson et al., 2002; Baldwin et al., 2003; Brönnimann et al., 2004; Cagnazzo and Manzini, 2009).

The seasonal forecasting capability has improved over time (Kharin and Zwiers, 2003; Palmer, 2004; George and Sutton, 2006; Stocker et al., 2013), especially in the Tropics. Some of the improvement can be attributed to more advanced models, but part of the progress can also be linked to better model initialisation (Zhang and Frederiksen, 2003; Li et al., 2008) or a better understanding of phenomena such as teleconnections (Hawkins et al., 2002). Recent seasonal forecasting activity has also included prognoses for storm statistics (Compo and Sardeshmukh, 2004; Vitard et al., 2007).

Murphy et al. (2010) reviewed the state on predicting decadal climate variability and change. They provided an

overview of phenomena with decadal time scales, such as the NAO, the Atlantic Multi-decadal Oscillation (AMO), the Tropical Atlantic surface air temperature Gradient Oscillation (TAG), the North Pacific Oscillation (NPO), the Pacific Decadal Oscillation (PDO), and the Inter-decadal Pacific Oscillation (IPO). The state-of-the-art models are able to simulate key features of the most prominent of these modes, they argued, and at least on a global scale, simulated decadal temperature variability has been shown to be realistic. They also noted that precipitation variance in many latitude bands may be underestimated by climate models.

There have been some studies of the skill of downscaled climate data in the context of seasonal-to-decadal predictions. For instance, Gershunov and Barnett (1998) analysed the ENSO influence on intraseasonal extreme rainfall in the Contiguous United States. They looked at both observations and model results and found an ENSO response in the frequency of occurrence of heavy rainfall in the Southeast, Gulf Coast, central Rockies, and the general area of the Mississippi–Ohio River valleys. Feddersen (2003) studied the predictability of seasonal precipitation in the Nordic region, and Feddersen and Andersen (2005) applied a statistical downscaling method to seasonal ensemble forecasts. Landman et al. (2009) examined the performance of dynamical and empirical downscaling for South Africa from seasonal forecasts, and Reason et al. (2006) studied the effects of the Atlantic Ocean state on subsequent seasonal-to-decadal Southern African weather statistics.

We wanted to test whether it is possible to predict local precipitation statistics on seasonal-to-decadal time scales, given a good forecast for surface air temperature (SAT) and mean SLP. Here we use the terms ‘forecast’ and ‘prediction’ as synonyms. A dependency of the local precipitation statistics to large-scale SAT and SLP conditions implies a kind of downscaling, and it is possible to use empirical-statistical downscaling (ESD) methods to analyse observations for such inter-dependencies. We also assume that some large-scale features such as the North-Atlantic SAT and SLP can be predicted by coupled GCMs on a lead time from seasons to decades, due to the persistence and inertia in the world oceans. This study can be described as a ‘perfect predictor experiment’ (Maraun et al., 2015).

Seasonal-to-decadal forecasts can be made for a number of different aggregated statistics, such as mean, standard deviation, frequencies, number of events, as well as parameters for the 24-hour distributions. The exponential distribution has been used to describe temporal variations in the precipitation characteristics (Wilks, 1998; Burgueo et al., 2005; Benestad, 2013) although the observations suggest that 24-hour precipitation has a heavy tail bias (Benestad et al., 2012a, 2012b). The heavy tail bias with respect to the exponential distribution appears to be site

specific and not time dependent. It has been more common to use the gamma rather than the exponential distribution to model 24-hour precipitation (Thom, 1958; Buishand, 1978; Wilks and Wilby, 1999; Wilby and Wigley, 2002). The gamma distribution is defined by two parameters: the shape and the scale parameters and is more appropriate for the analysis of the entire record of both wet and dry days. In this case, only the wet days were used to fit the exponential distribution. The mean associated with the gamma distribution is calculated as the product of the scale and shape parameters, whilst the exponential distribution is fully defined by the mean, which makes it more attractive.

We are not trying to analyse the extremes or estimate the return values in this paper. Extreme precipitation is normally analysed using Extreme Value Theory (EVT) that use the tail of the distribution for parameter fitting (Coles, 1999), however, return values derived from EVT usually assume a stationary probability density function (pdf). On the contrary, we assume that the pdf varies over time. The probability for precipitation amounts exceeding a given threshold is inferred from this assumption and the change in the entire pdf (the probabilities must add up to one).

2. Overview: methods & data

The simplest form of seasonal forecast for precipitation is a prediction of a total amount for a given season, or an average \bar{x} . However, precipitation has a number of different aspects, such as how often it rains, how much it rains on a rainy day, how long it is between each time it rains, and the duration of the rain event. The precipitation statistics differ between days when it rains (wet-day) and when it doesn't (dry day) because different physical processes and conditions are present. A complete picture of precipitation can be derived from parameters μ , f_w , the number of consecutive dry days (n_{cdd}), the number of consecutive wet days (n_{cwd}), type (rain, snow, hail), and the spatial extent; however, in this case we will limit the analysis to μ and f_w . The question is whether there are systematic dependencies between such phenomena and large-scale conditions such as particular SAT anomalies or SLP patterns. An estimate for the *wet-day* mean μ can be related directly to extremes and higher quantiles if the distribution of the amounts is close to being exponential (Benestad et al., 2012a, 2012b; Benestad, 2013). In this case, μ can be regarded as a proxy for the precipitation amount. Extreme precipitation events can lead to floods caused by long wet spells as well as droughts which are products of extreme long dry intervals.

One practical problem is that the number of valid data points in a rain gauge record depends on the number of wet-days, and a wet-day frequency of ~ 0.3 implies approximately 30 valid data points for each season. This limitation is also a concern for monthly totals because data

samples with a size of 30 (assuming it rains every day of a month) are low in terms of the central limit theorem (Wheelan, 2014) and the sample mean is subject to substantial statistical fluctuations. Furthermore, the stochastic nature of the sampling (a random point in space and a random interval in time) will be more pronounced for the higher quantiles and extremes. One solution could therefore be to calibrate statistical models on a larger data samples, such as using all days in a year for each estimate of μ and mean values \bar{x} so that a wet-day frequency in the range 0.25–0.30 yields approximately 100 valid data points (number of rain events). Another option could be to combine a given season from several consecutive years. One science question therefore is whether a statistical model for precipitation statistics is sensitive to the sample size. In other words, is a model calibrated on statistics derived from annual samples (365 days) valid for shorter and longer samples? The answer to this question would depend on whether there are different physical processes present that act on different time scales.

2.1. Data

The predictor data was taken from the 20th century re-analysis (Compo et al., 2011), based on monthly mean data and the ensemble mean. The reason for using this re-analysis was its long time span (1871–2012) needed for the evaluation of the method. We used the ensemble mean of monthly mean SAT (0.995 sigma level) and mean SLP. The spatial domain was set to 50°W–20°E/40°N–72°N; however, five other tests with slightly different domains were also carried out (not shown). Furthermore, a slight variation of the spatial domain was also used for stations other than Bjørnholt, and the results shown here used a domain that extended 60°W to 10°E and 20°S to 20°N from the location of the respective station. These tests represented a crude and simple strategy for examining the sensitivity of the results to a subjective configuration choice and provided a means for identifying a skilful set-up. It is important to keep in mind the fact that sampling different set-ups constitutes a selection ('cherry picking'), and field significance should be carried out for hypothesis testing (Wilks, 2006). However, this was only true for Bjørnholt and not the other sites. Furthermore, our strategy did not constitute a proper optimisation of the method.

The two annual mean parameters describing the wet-day mean μ and frequency f_w were derived from daily rain gauge records taken from the climate archives of MET Norway (<http://eklima.met.no>) and from ECA&D (Klein Tank et al., 2002). The rain gauge records were selected based on the criteria that they included at least 130 yr, as long records were needed for the split-sample test. Figure 1 shows the location of the rain gauges. The 24-hour

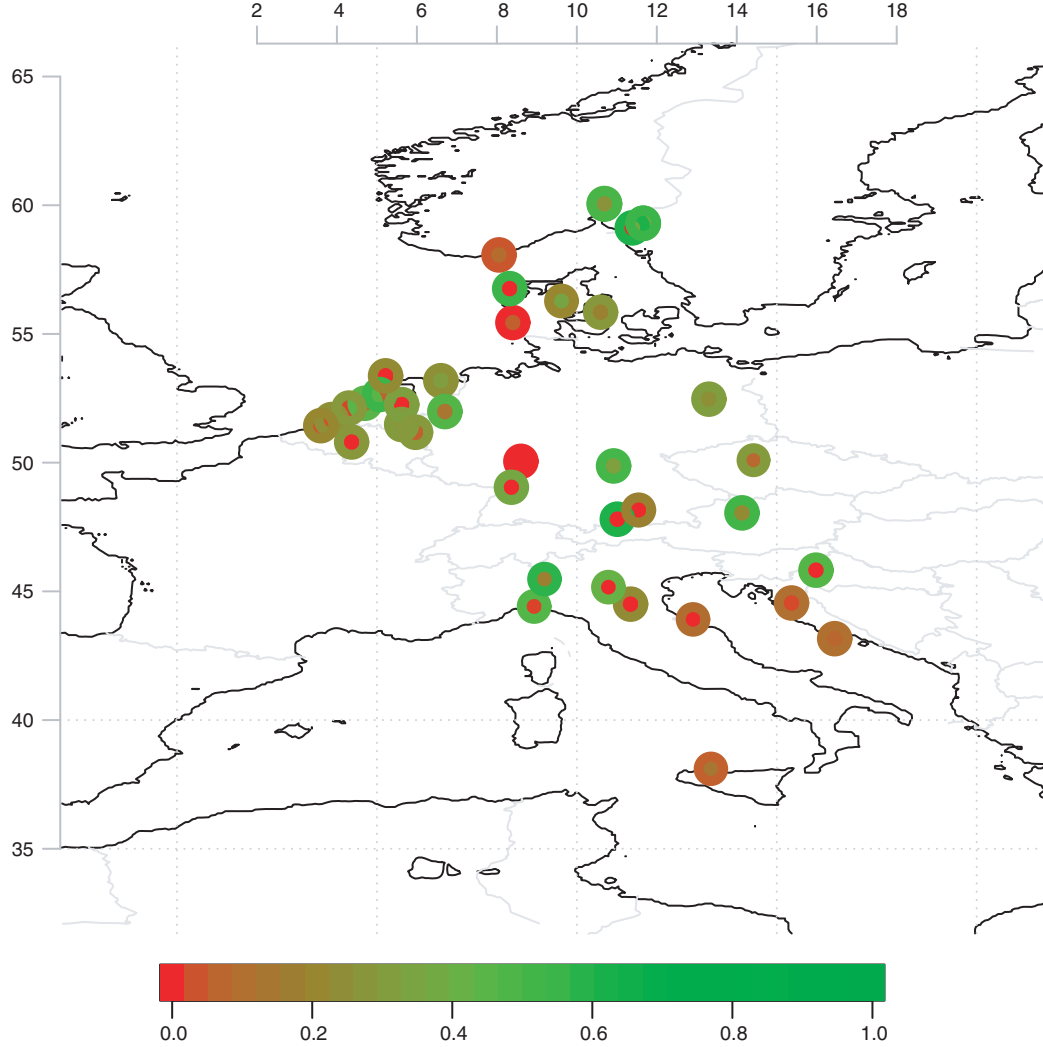


Fig. 1. The location of the rain gauge measurements for more than 130 yr. The size of the symbols is proportional to the number of valid data and their colour is according to the correlation between observed and predicted μ (inner part) or f_w (outer part) over the independent period.

precipitation record from Bjørnholt in the forest north of Oslo was used as the primary data record. We let the capital letter X represent any random 24-hour precipitation, whereas the lower case symbol x is used to refer to a specific value. The exact definitions of μ and f_w are provided in the Appendix.

2.2. Experiment design

Empirical orthogonal functions (EOFs) (Lorenz, 1956; North et al., 1982; Wilks, 1995; Benestad et al., 2008) can be used to represent the large-scale SAT and SLP conditions, here referred to as *predictors*. A third predictor e_s was derived from the SAT to represent the saturation vapour pressure according to the Clapeyron–Clausius equation

(saturation evaporation pressure $e_s = 10^{(11.40 - 2353/T)}$ where T is temperature in degree Kelvin). The principal components (Strang, 1988) describe how the presence of the different large-scale EOF patterns vary over time, and were used as the independent variables (covariates) in an ordinary linear multiple regression. The regression involved a step-wise screening, based on the Akaike information criterion (Akaike, 1974). The regression-based ESD assumed that the data followed a normal distribution (see the Appendix).

The dependent variable in the regression represented the local precipitation statistics (μ and f_w), and is referred to as *predictand*. Both predictors and predictands were aggregated on an annual basis for the calibration of the down-scaling model. Both μ and f_w can be regarded as an estimate

of the mean that, according to the central limit theorem, converges to a normal distribution with sample size (f_w is the mean of a series of zeroes and ones). We also checked how well μ and f_w conformed with the normal distribution through an analysis of the residuals. The precipitation data and the EOFs were also aggregated into seasonal and 5-yr (pentad) statistics respectively, to test whether a model trained on annual samples provided skilful predictions for smaller and larger samples. The purpose was to test the general validity of the model.

Common EOFs (Barnett, 1999; Benestad, 2001) were used for the model calibration to ensure that the seasonal and pentad EOFs corresponded to the same spatial patterns as the annual EOFs. The EOFs were computed independently for the large-scale atmospheric flow (SLP) and temperature anomalies (SAT), and the predictions were based on a step-wise approach where the ESD model first was calibrated on the e_s EOFs for the downscaling of the wet-day mean μ and the SAT EOFs for f_w respectively. The e_s or SAT-based regression model was then used to predict the local quantities, and the resulting residuals were then used to calibrate another regression model that used SLP EOFs as predictors. The final predictions based on both e_s /SAT and SLP were then synthesised by adding the predictions from the two steps. Limiting the EOFs to the first six leading ones imply both a selection of the largest spatial scales and reduces the risk of over-fit (Wilks, 1995), however, 12 EOFs were used for the locations other than Bjørnholt in order to improve the chances of finding a connection between local and large scales with a crude specification of predictor domain. A third step was introduced for the pentads to account for a possible link to the global mean temperature on the longer time scales (Benestad, 2013), where the residual of the downscaled pentad μ from the SAT and SLP EOFs was subject to a final regression against the global mean temperature as in Benestad (2013).

The experiments and analyses were designed for the evaluation of forecasts against independent data, and hence the ESD model used the interval 1900–1970 for calibration and independent data outside this interval for independent (out-of-sample) evaluation. The analysis was implemented in the R system for statistical computing (R Development Core Team, 2008, <http://www.R-project.org/>) and used the new R-package ‘esd’ (see the Appendix for more details).

2.3. Validation

The skill of the forecasts was assessed through the estimation of the Pearson correlation in the predicted and observed annual values μ and f_w . One reason for analysing annual mean values for these quantities is the expected characteristics of stochastic data of similar type: a normal

distribution for the average of a continuous variable. Likewise, the skill of the predicted values for the annual wet-day 90th percentile q_{90} was assessed in terms of its correlation against corresponding quantity from the observations. The skill of predicted probabilities was assessed by comparing the annual observed number of events with precipitation exceeding a high threshold (10 mm/day) against a predicted 90% confidence interval, based on probability estimates and a binomial distribution. One diagnostic involved comparing the residual to Gaussian white noise (distribution shape, trend, and lagged correlation).

3. Results

Figure 2 shows the downscaling results for the annual mean μ at Bjørnholt, which was the same data record as Benestad and Melsom (2002) found a teleconnection link in the North Atlantic. The upper part of the figure shows the observations and predictions. For the independent data, the correlation was 0.27 (with 90% confidence interval of 0.01–0.49), and for the entire period, the correlation was 0.38 (0.22–0.52). The predicted values in the calibration interval are shown in faint colours, and the residual is shown in the lower part of the figure. The inserted diagrams show the auto-correlation function (ACF) and the normal quantile–quantile plot (q–q normal) for the residuals. One indication that the analysis has identified all dependencies is when the residuals have low auto-correlation and are close to being normally distributed (white noise). Hence, these results suggest that the downscaling accounts for most of the connection to larger spatial scales. Table 1 (first column) shows similar scores for some European stations whose locations are shown in Fig. 1, where the colour scale of the inner part of the symbols indicates the correlations for μ . For a number of stations, the evaluation indicated no skill, however, there were also some stations with higher correlations, the highest being 0.42 for Strømsfoss sluse. Both the estimated and downscaled μ for Bjørnholt suggest a slight increase during the last evaluation period since 1960.

Figure 3 shows the results from a similar analysis as in Fig. 1, but for f_w rather than μ . The correlation between predicted and observed annual wet-day frequency was 0.48 for the out-of-sample data, with a 90% confidence interval of 0.26–0.66. Hence the results were statistically significant at the 5%-level. The ACF gave no indication of persistence in the residual and the qq-plot showed that the residuals were close to being normally distributed. However, the residual had a number of pronounced jumps (‘spikes’) between intervals where the values varied to a much smaller degree and with a somewhat more persistent character. Again, the inspection of the residual suggests that the downscaling captured most of the large-scale dependency in f_w .

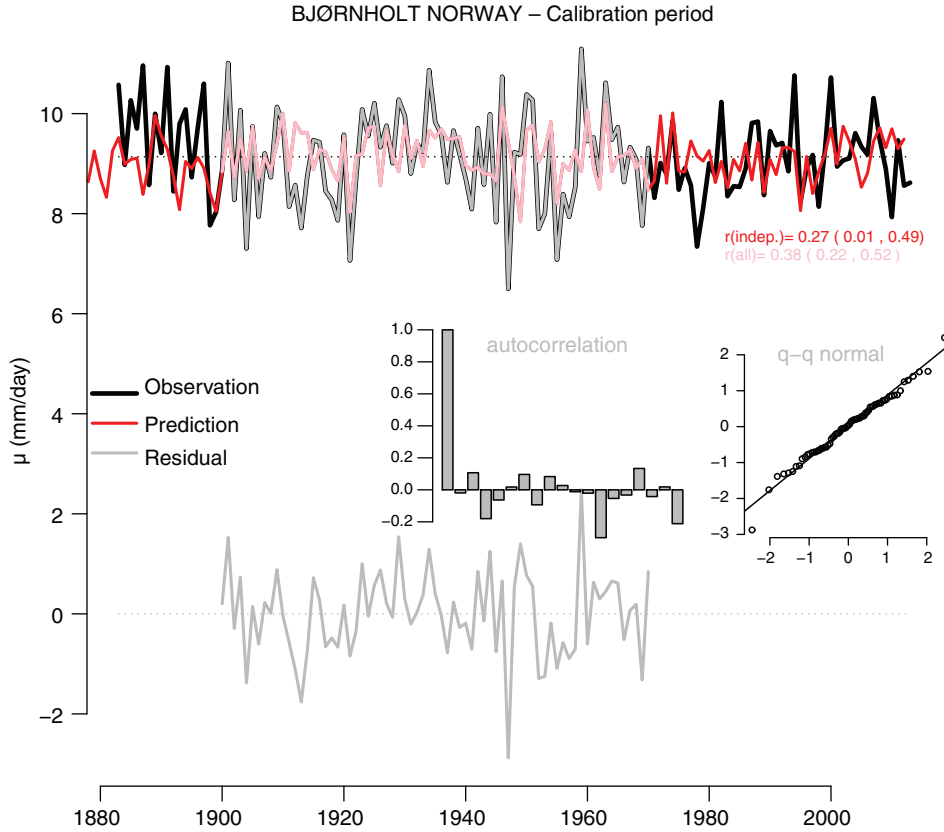


Fig. 2. An assessment of the downscaled μ . The upper part shows μ estimated from the rain gauge data (black) and predicted through the downscaling (red). The faint sections of the curves indicate the calibration interval. The correlation for the independent sections of the data are 0.27 the out-of-sample years and 0.38 for the entire curve, both statistically significant at the 5%-level. The lower part of the figure shows the residuals in grey, and the inserts show the auto-correlation function (ACF) and a q-q plot of the residuals against a normal distribution. The ACF indicates no persistence and the q-q plot suggest that the residuals follow the normal distribution closely.

Since 1970, both the estimated and downscaled f_w suggest a slight increase, in other words, there has been more rainy days. The first column in Table 2 indicates that the correlations for f_w were in general higher than for μ , often exceeding 0.50. The skill scores for f_w are indicated by the colour of the outer circle of the symbols in Fig. 1. The highest score was 0.63 for Halden and Hohenpeissenberg.

Figure 4 shows a set of maps indicating the regional patterns in surface temperature (upper) and mean sea-level pressure (lower) associated with high values for μ (left) and f_w (right) at Bjørnholt. The teleconnection patterns associated with μ exhibited a temperature gradient with cold conditions over the British Isles and the Bay of Biscay and warm conditions over northern Norway. The corresponding pattern for the SLP was characterised by low pressure to the west of the British Isles. However, the patterns associated with high frequency of wet days differed from those related to high amounts: cold conditions south of Greenland and west of Britain and warm over the central–northern European continent.

Figures 2–4 presented results for annual statistics. One interesting question is whether models calibrated on annual samples of μ can be used to predict variations on seasonal time scales. In this case, the definition of ‘independent’ was the seasons for the years which were not used for the calibration, and category ‘all’ also included the seasons for those years on which the annual aggregates were calibrated. Here we report the skills for all years, since the seasonal aggregates for the calibration period were not used to train the model. Figure 5 shows the results of such a test, however, the predictions for the seasons had little skill for μ during winter (0.15), but higher correlations for spring, summer and autumn (0.25, 0.29, 0.39, respectively; Table 1). Seasonal differences have also been noted in earlier studies (Wilby and Wigley, 2000; Hanssen-Bauer et al., 2003). One difference between the winter and the other seasons is that more of the precipitation is in the form of snow. The predicted values were slightly biased with respect to the observations and did not capture the peaks, but there were only minor differences between the mean levels during

Table 1. Summary of correlation scores: Annual (both for the independent out-of-sample and for all years), seasonal and pentads of wet-day mean (μ) of precipitation

	Annual (ind.)	Annual (all)	q_{90}	DJF	MAM	JJA	SON	Pentads
BJØRNHOLT	0.27	0.38	0.28	0.15	0.25	0.29	0.39	0.18
BAMBERG	0.33	0.41	0.30	0.11	0.16	0.23	0.13	0.57
BERLIN-DAHLEM	0.25	0.25	0.17	0.23	−0.08	0.14	0.06	0.33
BOLOGNA	−0.02	0.07	0.11	0.11	0.04	−0.04	−0.13	−0.29
FRANKFURT	−0.04	0.15	0.12	0.05	0.00	0.09	0.07	0.29
GENOA	0.01	0.24	0.27	0.24	0.07	−0.07	0.13	−0.05
GOSPIC	0.02	0.08	0.10	0.01	0.03	−0.02	−0.08	0.37
GRONBAEK-ALLINGSKOVGARD	0.35	0.21	0.18	0.01	0.13	0.10	0.01	0.26
GRONINGEN-1	0.32	0.52	0.42	0.08	0.09	0.43	0.13	0.76
HALDEN	0.00	0.21	0.13	0.38	0.18	0.08	0.12	−0.07
HELMOND	0.28	0.39	0.32	0.08	0.17	0.29	0.19	0.25
HOHENPEISSENBERG	−0.06	0.09	0.15	−0.06	0.00	0.02	0.03	−0.19
HOOFDDORP	0.13	0.12	0.09	0.13	0.08	0.27	0.09	0.06
HOORN	0.06	0.21	0.15	0.20	0.11	0.33	0.15	0.19
HVAR	0.07	0.09	0.01	0.11	0.05	0.06	−0.07	NA
KARLSRUHE	−0.04	0.10	0.11	−0.07	0.14	0.06	0.08	0.14
KERKWERVE	0.32	0.35	0.33	0.16	0.02	0.32	0.18	0.62
KREMSMUNSTER	0.23	0.40	0.28	0.13	0.27	0.25	0.09	0.11
MANTOVA	−0.27	0.07	0.06	0.02	0.11	0.11	0.20	NA
MILAN	0.18	0.36	0.30	0.04	0.04	0.18	0.27	NA
MUENCHEN	−0.08	0.02	0.00	0.11	0.10	−0.09	−0.07	−0.35
NORDBY (FANO)	0.06	0.04	−0.06	0.18	0.13	−0.04	0.16	0.03
OKSOEY FYR	0.09	0.26	0.24	0.26	0.01	0.15	−0.10	0.34
PALERMO	0.14	0.41	0.32	0.24	0.15	−0.05	0.21	NA
PESARO	−0.12	0.06	0.04	−0.09	0.07	−0.05	0.13	NA
PRAHA-KLEMENTINUM	0.08	0.17	0.04	−0.23	−0.11	0.07	0.09	NA
PUTTEN (GLD)	−0.12	0.26	0.29	0.12	0.15	0.33	0.06	0.43
ROERMOND	0.06	0.28	0.29	0.01	0.16	0.22	0.12	0.35
SCHEVENINGEN	0.01	0.16	0.20	−0.15	0.11	0.16	0.02	0.08
STRØMSFOSS SLUSE	0.42	0.36	0.27	0.04	0.04	0.19	0.18	0.37
TRANEBJERG	0.19	0.40	0.31	0.36	0.18	0.24	0.04	0.52
UCCLE	−0.20	0.11	0.03	0.29	0.03	0.10	0.12	−0.12
VESTERVIG	−0.12	0.15	0.07	0.20	0.18	0.20	−0.02	−0.15
VLISSINGEN-1	0.04	0.18	0.16	0.09	−0.08	0.26	0.29	0.23
WEST TERSCHELLING	−0.09	0.37	0.26	0.33	0.15	0.25	0.00	0.30
WINTERSWIJK	0.14	0.38	0.30	0.13	0.11	−0.02	0.16	0.35
ZAGREB-GRIC	−0.02	0.14	0.15	0.17	0.08	−0.09	0.15	−0.18

The column ‘ q_{90} ’ provides the correlation scores for the predicted q_{90} as in Fig. 8. The correlations for the seasons and pentads were for all years.

the different seasons. Table 1 suggests that findings for Bjørnholt of generally low-to-moderate seasonal correlation skills for μ (columns 4–7) and high for f_w (Table 2) are also fairly typical for the other European stations. A similar analysis for f_w (Fig. 6) suggested even higher scores. Downscaling calibrated for the annual mean f_w at Bjørnholt gave a correlation of 0.67 for winter, 0.64 for spring, 0.55 for the summer, and 0.76 for the autumn season.

For the pentads (5-yr aggregated values), the correlation between predictions of μ and the wet-day mean derived from the observations was 0.18 for all years (−0.22, +0.25)

and not statistically significant (Fig. 7). It is worth bearing in mind that the number of data points in the correlation analysis was lower than for the annual analysis and makes a difference in the estimation of the confidence intervals. The ACF gave no indication of persistence in the residual, but rather a slight lag-one anti-correlation. The qq-plot for the residuals after the regression analysis against the (regional) air surface temperature and mean SLP showed a divergence from the diagonal in the lower tail. The lack of skill on the 5-yr timescale was fairly typical for those sites with sufficient record lengths (Table 2; last

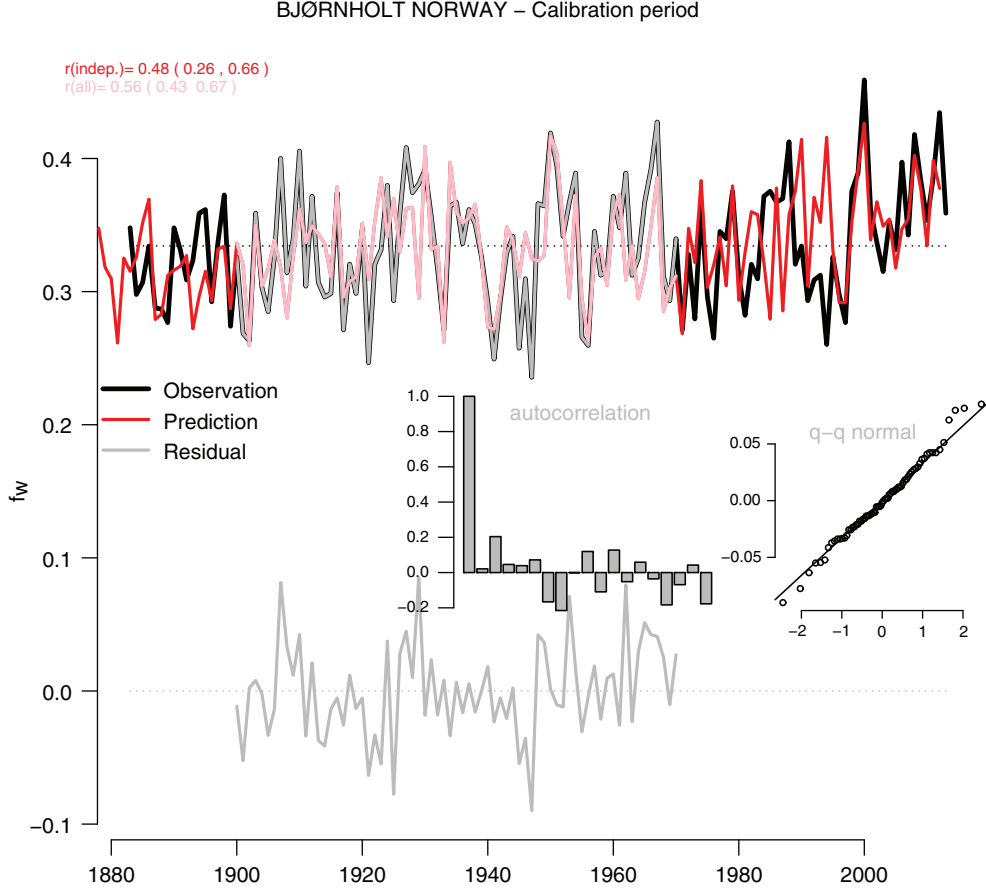


Fig. 3. Same as Fig. 2 but for f_w . The correlation between the downscaled results and observed wet-day frequency is 0.48 (0.56 for the entire record) and statistically significant at the 5%-level. The two curves follow closely, albeit with periods of poorer match such as in the 1890s and 1990s. The residuals do not indicate the presence of persistence and appear to follow the normal distribution closely. Similar correlations for independent and the entire record suggest robust results.

column). Similar exercise for f_w rather than μ indicated a reduction in the skill for the wet-day frequency on the 5-yr timescales (Table 2; last column).

Figure 8 shows a comparison between the annual wet-day 90th percentile q_{90} from the observations, the corresponding values estimated from μ assuming an exponential distribution, and \hat{q}_{90} predicted from SAT and SLP. The latter prediction took μ from Fig. 2 and solved for $\hat{q}_{90} = -\ln(1-p)\mu(\text{SAT}, \text{SLP})$. The year-to-year correlation between the downscaled and observed values was 0.28, and statistically significant at the 5%-level. In particular, the downscaling did not capture the magnitude of the major peaks in q_{90} ; however, the value of q_{90} derived directly from the observed μ gave a closer reproduction of the multiyear variations (grey; correlation of 0.76). The q_{90} was chosen rather than q_{95} , as the former was less influenced by statistical fluctuations ($f_w \sim 0.3$ gives about 120 data points per year) and the annual q_{95} was strongly affected by a single year. In Table 1, the corresponding

correlations are presented for the other European stations (column 3), and the general picture was a moderate and positive correlation between downscaled and observed q_{90} .

One way to use the downscaled results for forecasting is to estimate the number of events exceeding a certain threshold value each year. The method predicts the probability $Pr(X > x_0)$ based on the exponential distribution and f_w , and the confidence interval for the number of events can be estimated according to the binomial distribution (see the Appendix). For each year, a 90% confidence interval for the predicted number of events was estimated according to the predicted probability, which was compared with the observed number of events (henceforth ‘counts’). We would expect some of the counts to fall outside the confidence interval even if the predicted and observed data have an identical distribution, and the sum of the number of years out of a total of N years with the observed counts outside the predicted interval ($\hat{n}_{05}(t) - \hat{n}_{95}(t)$) is also expected to

Table 2. Summary of correlation scores. Annual, seasonal and pentads of wet-day frequency (f_w) of precipitation

	Annual (ind.)	Annual (all)	$n(X > 10 \text{ mm})$	DJF	MAM	JJA	SON	Pentads
BJØRNHOLT	0.48	0.56	0.59	0.67	0.64	0.55	0.76	0.61
BAMBERG	0.51	0.57	0.46	0.49	0.48	0.23	0.66	0.41
BERLIN-DAHLEM	0.31	0.33	0.23	0.33	0.28	0.24	0.38	0.18
BOLOGNA	0.23	0.46	0.28	0.47	0.37	0.14	0.42	0.65
FRANKFURT	-0.10	0.18	0.16	0.13	0.10	0.07	0.51	0.19
GENOA	0.48	0.63	0.28	0.55	0.47	0.26	0.72	0.44
GOSPIC	0.11	0.40	0.41	0.30	0.38	0.28	0.36	0.32
GRONBAEK-ALLINGSKOVGARD	0.21	0.01	0.04	0.34	0.00	-0.29	0.21	0.13
GRONINGEN-1	0.26	0.44	0.30	0.69	0.66	0.40	0.66	-0.05
HALDEN	0.63	0.60	0.44	0.40	0.58	0.37	0.72	0.09
HELMOND	0.30	0.47	0.40	0.67	0.55	0.37	0.66	-0.03
HOHENPEISSENBERG	0.63	0.65	0.16	0.63	0.44	0.32	0.72	0.31
HOOFDDORP	0.45	0.57	0.23	0.62	0.74	0.48	0.70	0.21
HOORN	0.55	0.62	0.28	0.77	0.63	0.46	0.66	0.42
HVAR	0.11	0.50	0.44	0.50	0.31	0.24	0.34	0.67
KARLSRUHE	0.36	0.40	0.26	0.48	0.39	0.40	0.60	0.27
KERKWERVE	0.25	0.36	0.35	0.66	0.59	0.40	0.67	-0.29
KREMSMUNSTER	0.52	0.63	0.33	0.64	0.58	0.36	0.60	0.39
MANTOVA	0.42	0.57	0.20	0.48	0.51	0.13	0.52	NA
MILAN	0.59	0.71	0.44	0.52	0.54	0.35	0.62	NA
MUENCHEN	0.19	0.30	0.10	0.47	0.29	0.31	0.46	0.22
NORDBY (FANO)	-0.06	-0.07	0.10	-0.41	-0.14	-0.16	-0.06	-0.03
OKSOEY FYR	0.04	0.33	0.20	0.16	0.47	0.38	0.29	0.33
PALERMO	0.06	0.20	0.21	0.35	0.23	0.26	0.24	NA
PESARO	0.10	0.32	0.26	0.52	0.33	0.13	0.40	NA
PRAHA-KLEMENTINUM	0.30	0.27	0.02	0.13	0.24	0.14	0.15	NA
PUTTEN (GLD)	0.32	0.48	0.34	0.62	0.55	0.51	0.65	0.14
ROERMOND	0.29	0.41	0.28	0.55	0.52	0.30	0.66	-0.08
SCHEVENINGEN	0.30	0.49	0.35	0.61	0.55	0.47	0.66	0.11
STRØMSFOSS SLUSE	0.41	0.52	0.44	0.54	0.57	0.46	0.68	0.16
TRANEBJERG	0.29	0.50	0.35	0.62	0.54	0.60	0.64	0.39
UCCLE	0.28	0.56	0.29	0.61	0.56	0.61	0.74	0.24
VESTERVIG	0.56	0.60	0.30	0.76	0.63	0.60	0.79	-0.05
VLISSINGEN-1	0.21	0.38	0.21	0.49	0.59	0.48	0.62	-0.13
WEST TERSCHELLING	0.23	0.41	0.34	0.65	0.60	0.57	0.69	-0.10
WINTERSWIJK	0.46	0.53	0.38	0.67	0.59	0.42	0.64	-0.07
ZAGREB-GRIC	0.47	0.60	0.40	0.57	0.36	0.28	0.59	0.62

The column ' $n(X > 10 \text{ mm})$ ' shows the correlation between the median number of days with rainfall greater than 10 mm/day as in Fig. 9. There was a bias for some stations that gave a shift in the predicted 90% confidence interval even with moderately high correlations. The correlations for the seasons and pentads were for all years.

follow a binomial distribution for perfect predictions. Figure 9 shows a comparison between predicted number of precipitation events exceeding 10 mm/day [a threshold used by road authorities in Norway (Frauenfelder et al., 2013)] and the counts of observed events. The downscaled results reproduced the realistic multiyear variability with a correlation between the predicted annual median count $n_m(t)$ and observed count $n_o(t)$ of 0.6. However, the probability that the predictions conformed to the same statistical distribution as the 130 yr record was only 1% (i.e. different at the 5% statistical significance level). The

predicted counts were outside the confidence interval in 22 of 130 yr, compared to a 90% confidence interval of 8–19 according to the binomial distribution. In other words, the predicted annual count of annual heavy-precipitation events based on the downscaling exhibited a wider distribution than the observed statistics with a high proportion of values outside the 90% confidence interval ($p = 1\%$). On the other hand, the same analysis applied to the annual wet-day means and frequencies taken directly from the observations suggested a narrower distribution and gave only two cases with counts outside the 90% confidence

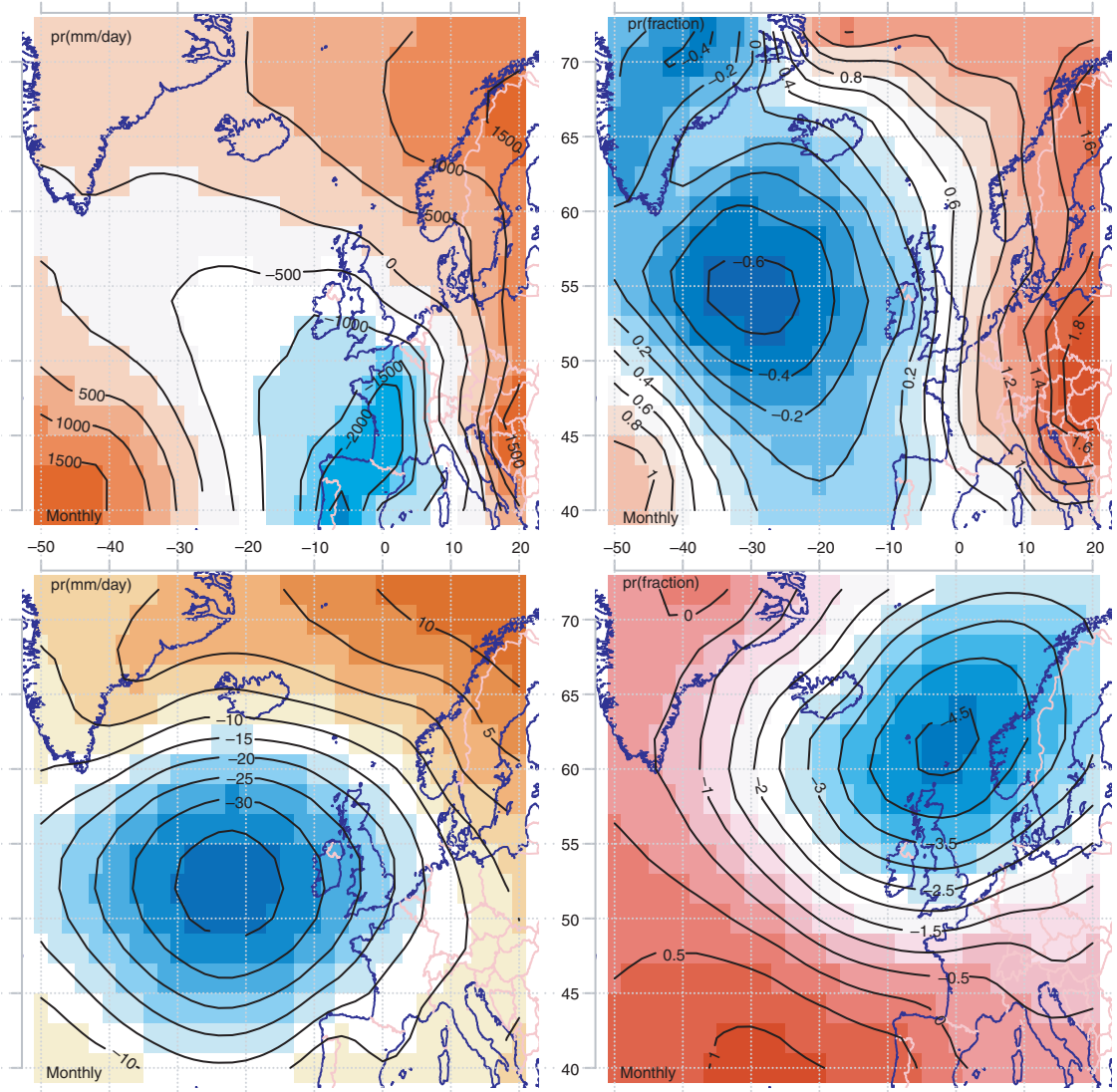


Fig. 4. Maps showing the spatial SAT (upper) and SLP (lower) anomalies associated with variations in μ (left) and f_w (right).

interval, significantly lower than expected (8–19). The assessment of skill should span over a range of threshold values, as shown in Fig. 10, displaying the probability of N being anomalously high or low (thin lines with symbols) for a range of different thresholds (x-axis). For thresholds x_0 in the range 20–75 mm/day, the predicted results suggested a good correspondence with a binomial distribution with $p=0.1$. The number of observed counts outside the predicted 90% confidence interval (axis on the right), however, suggests that the value estimated from the observed annual mean μ and f_w was almost always within this range, and hence, the estimated confidence intervals were too wide (the predictions were ‘under-confident’), with the exception of low threshold values ($x_0 < 10$ mm/day).

4. Discussion

The intention here was to emphasise the wet-day mean, as it is related to the statistics for variables that approximately follow an exponential distribution. Another point was to link the downscaled wet-day mean and frequency to the higher-order statistics, such as quantiles, number of days with heavy precipitation, and probabilities of exceedance.

One explanation why the correlation skill for q_{90} was lower than for $n(X > 10 \text{ mm/day})$ may be that the latter was derived from both f_w and μ , that $\Pr(X > x_0) \propto f_w$, and that the downscaling of f_w was associated with fairly good skill. The estimate of μ also provided a reasonable approximation for the shape of the pdf on which the results were based. A simple test was carried out where either μ or f_w

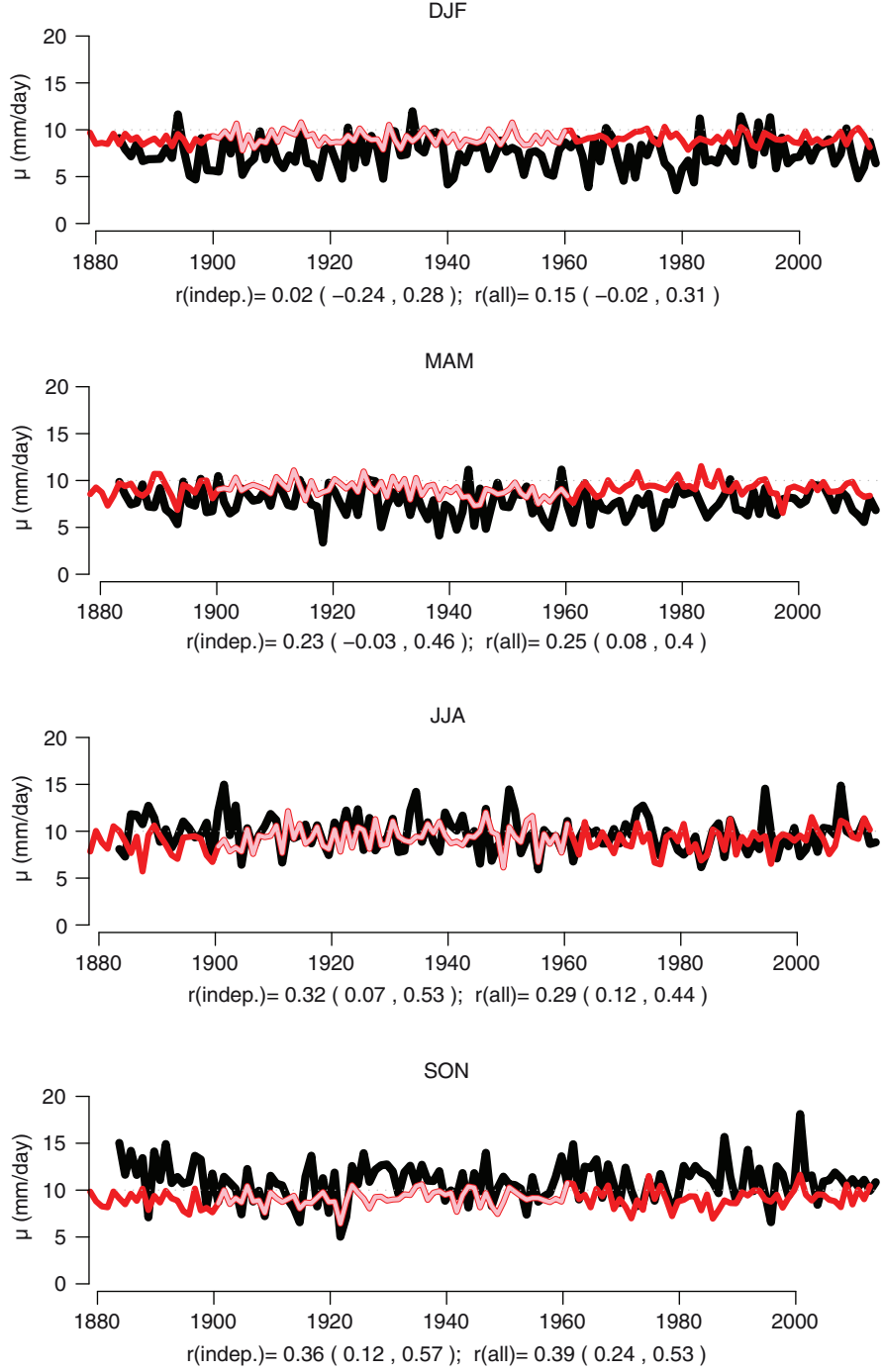


Fig. 5. A comparison between seasonal μ for the four seasons and corresponding values predicted through the downscaling (red). The section of curves with faint colours indicate the calibration period. The correlation is distinguishable from zero only in summer. The results do not indicate that a model calibrated on annual data is valid for sub-annual time scales. Similar correlations for independent and the entire record suggest robust results.

were constant and set to their mean value respectively (see the supporting material). The results of this test suggested that both μ and f_w contribute to the estimate of the probability, but that f_w has a stronger influence on the results.

Another interesting question is why μ seems to be more difficult to predict than f_w . It was in general more difficult to establish a link between the large-scale climate conditions and local μ , and the skill scores for μ were low for a

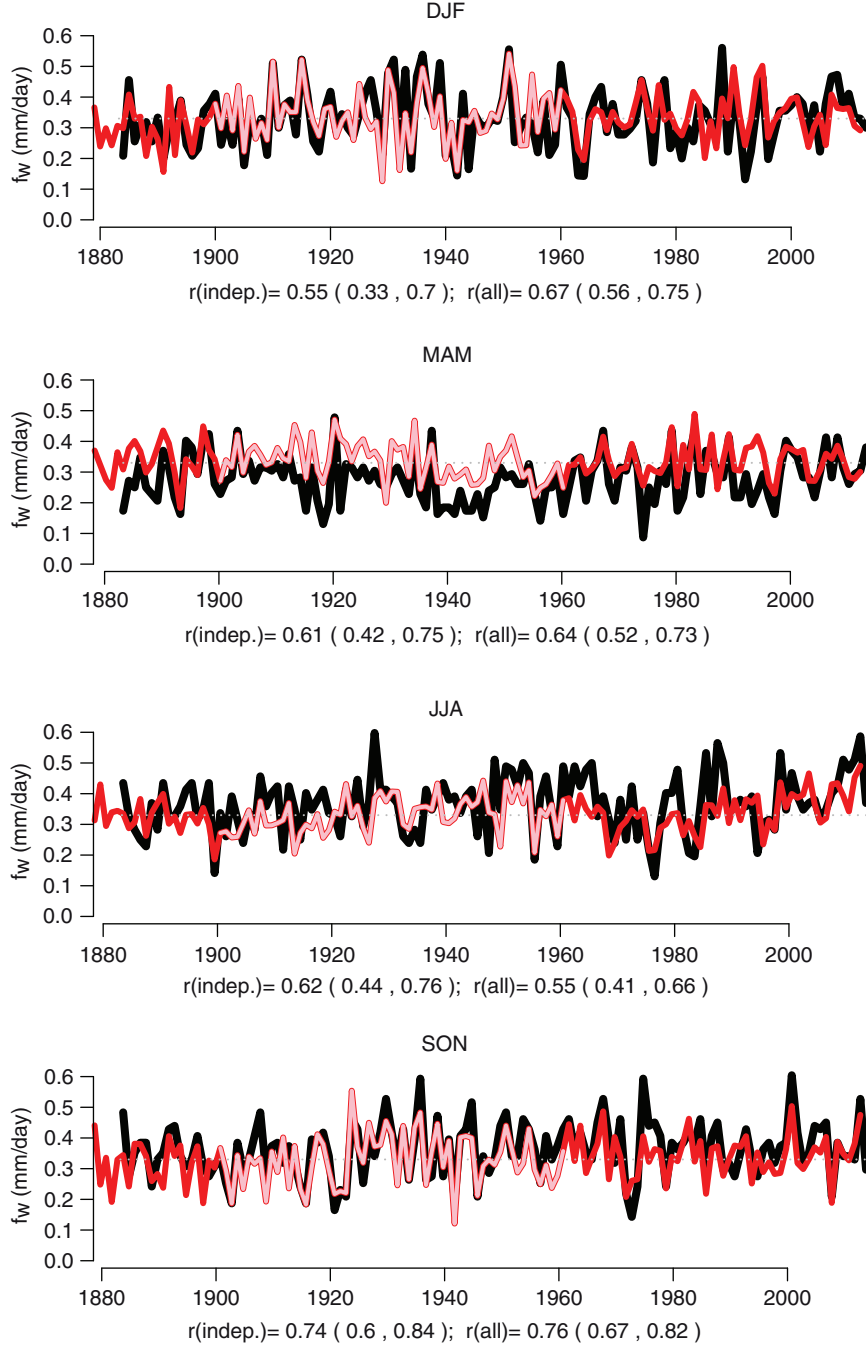


Fig. 6. A comparison between seasonal f_w for the four seasons and corresponding values predicted through the downscaling (red). The section of curves with faint colours indicate the calibration period. The correlation is distinguishable from zero only in summer. The results indicate that a model calibrated on annual data is valid for sub-annual time scales. Similar correlations for independent and the entire record suggest robust results.

number of locations in Europe despite generally good results for f_w . Figure 1 showed no clear geographical pattern regarding the skill scores for μ , which is consistent with a high sensitivity to local conditions or a complex dependency where the amounts depend on several factors

not accounted for. A principal component analysis (PCA) of annual variations in μ supports the notion of a weak dependency to large-scale conditions across the European stations (Benestad, 2013), as does the contrast with a PCA on annual f_w (supporting material). A canonical correlation

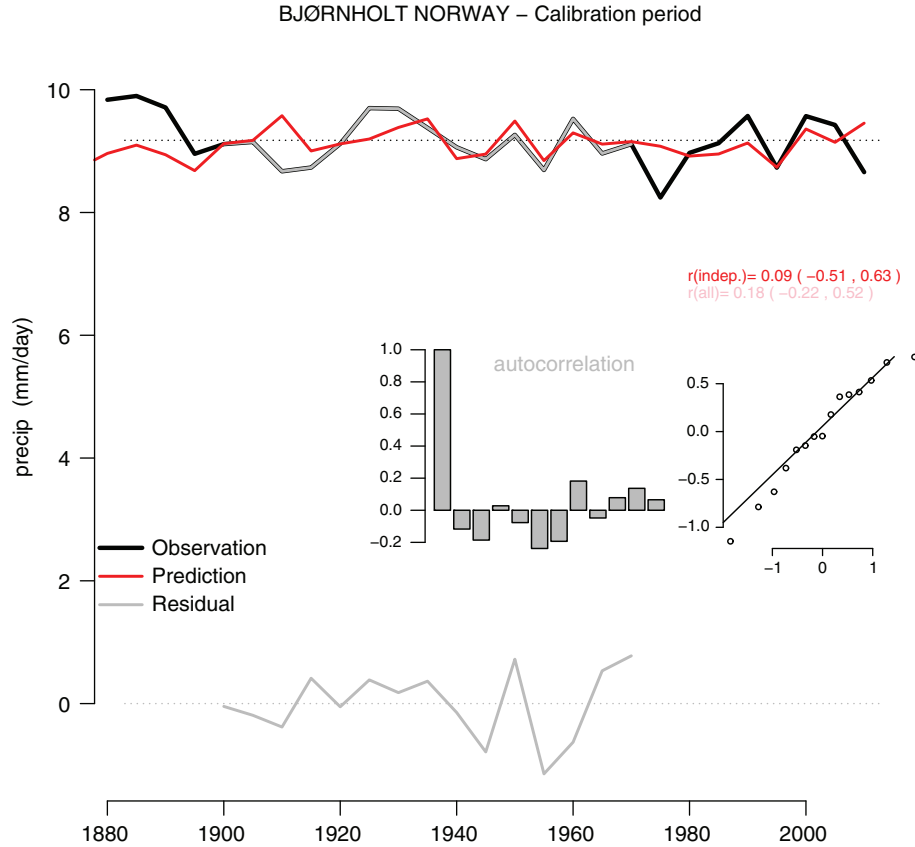


Fig. 7. Same as Fig. 2 but for pentads. The residual does not suggest any persistence and is close to being normally distributed, after a third regression against the global mean temperature was included. There is a modest correlation between the predictions and observations, but the results do not indicate that a model calibrated on annual data is valid for multi-annual time scales.

analysis (Wilks, 1995) also suggests a more homogeneous spatial pattern in f_w than in μ . The scores for f_w in Fig. 1 also suggest generally good results in the north and reduced skill in the south and around the Mediterranean. Figure A2 in the supporting material shows a similar analysis as in Fig. 1 but for all years, and the inclusion of the dependent data clearly improves the scores. Both non-stationarity and varying data quality with time (Wijngaard et al., 2003) may explain some of these differences.

The physical processes and conditions involved differ for μ and f_w (Linderson et al., 2004), where μ is expected to bear the signature of the cloud microphysics and may be more strongly influenced by summer-time convection and associated temporal and spatial scales. A plausible physical picture is that the connection to large-scales involves mechanisms through which humidity is advected with a geostrophic flow caused by pressure differences. Another factor is the dependency of the rate of evaporation on the surface temperatures, with increased evaporation due to higher temperatures in accordance to the Clapeyron–Clausius equation. In addition, large-scale temperature gradients can

affect the baroclinic stability and hence the genesis as well as the course of low-pressure systems, often associated with precipitation.

If higher μ in summer were due to more convective events, then the results for Bjørnholt may suggest that it may to some extent be possible to predict the aggregated values for μ from large-scale temperatures, vertical stability, or humidity. Another question is whether the low skill for μ for Bjørnholt in winter can be attributed to more snow and less liquid rain, for instance under-catch (Wolff et al., 2014). Another explanation is that low values of μ during winter means that the calibration with its annually aggregated value emphasises the warmer seasons with typically higher values. The predictor map for μ (Fig. 4; left panels) suggests that the higher values of the wet-day mean at Bjørnholt were associated with a low-pressure system over the British Isles and a southwest–northeast temperature gradient.

The corresponding results for the other stations gave a mixed picture, with some locations exhibiting similar scores as for Bjørnholt (green symbols in Fig. 1) whereas others

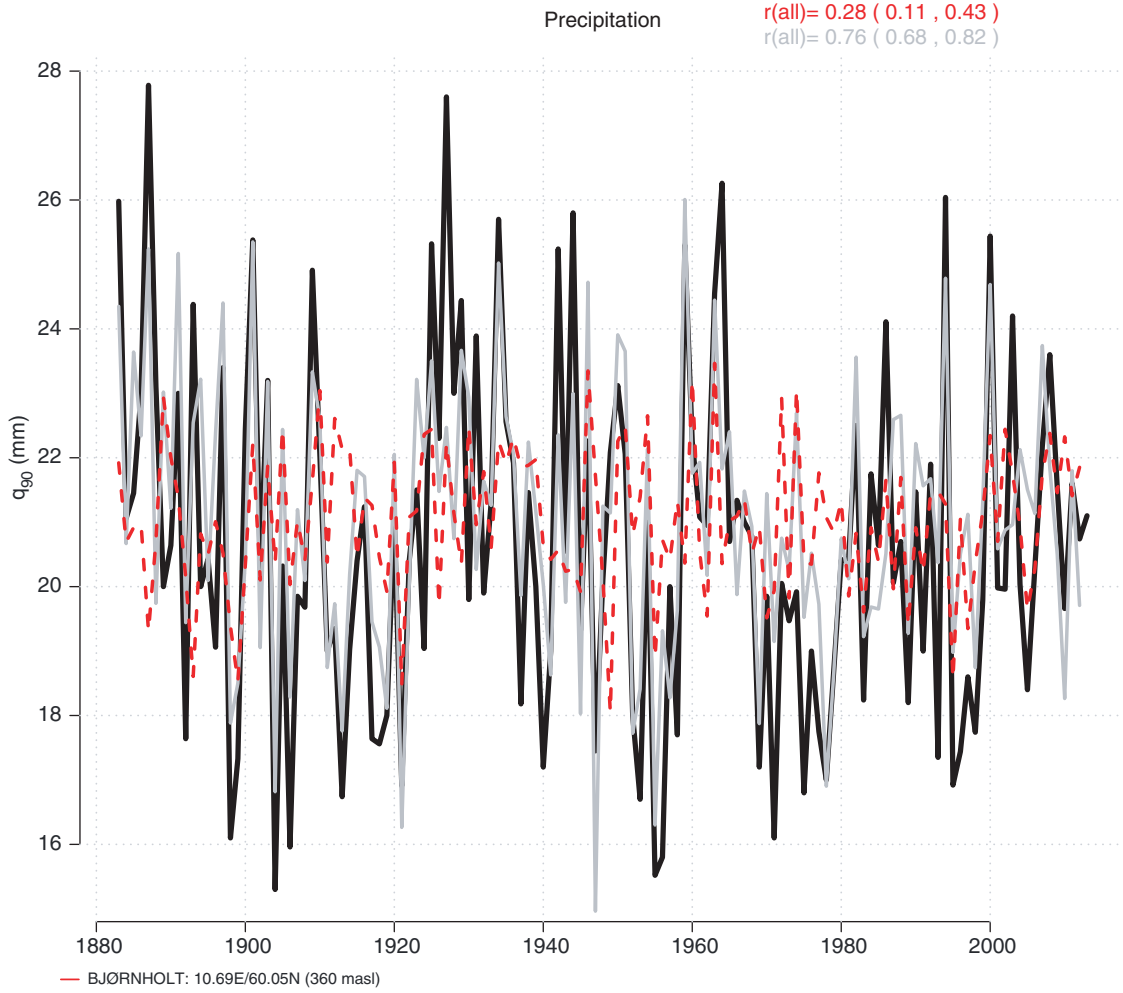


Fig. 8. A comparison between observed wet-day 95-percentile q_{95} and corresponding values estimated from μ (grey) and through downscaling (red) using SAT and SLP as predictors. The method appears to capture some of the multiyear variations but not some of the most pronounced individual spikes.

indicated low skill (red symbols in Fig. 1). There may be several reasons for the geographical variations in skill, such as the use of a fixed predictor domain size and the different locations being exposed to different prevailing winds and weather types. Another reason may be that the ECA&D data set is not homogeneous (Wijngaard et al., 2003). Furthermore, the experimental set-up may not have sufficiently captured the geographic variations in the link between small and large scales, and including 12 EOFs is associated with a higher risk for over-fit than including only six.

The marginal distribution for 24-hour precipitation tends to have a fat tail compared to that of the exponential distribution (Benestad et al., 2012a, 2012b), and hence the exponential distribution will tend to under-estimate the probabilities associated with extreme precipitation.

More accurate results are obtained with proper EVT (Coles, 1999). The Appendix provides a crude comparison between return-values derived based on EVT and the framework based on the exponential distribution presented in this paper. If the imprecision associated with the latter is acceptable, then crude estimates for daily precipitation return periods can be derived from downscaled results for μ and f_w .

It is also questionable whether the 20 C reanalysis used in these ‘perfect predictor experiments’ provide a sufficient representation of the past SAT and SLP. Changes in the surface network of observations may have an effect on its degree of homogeneity, and the lack of upper-air observations and imperfect estimates of historical sea surface temperatures imply that the reanalysis has some limitations. We have not addressed those here, but such caveats

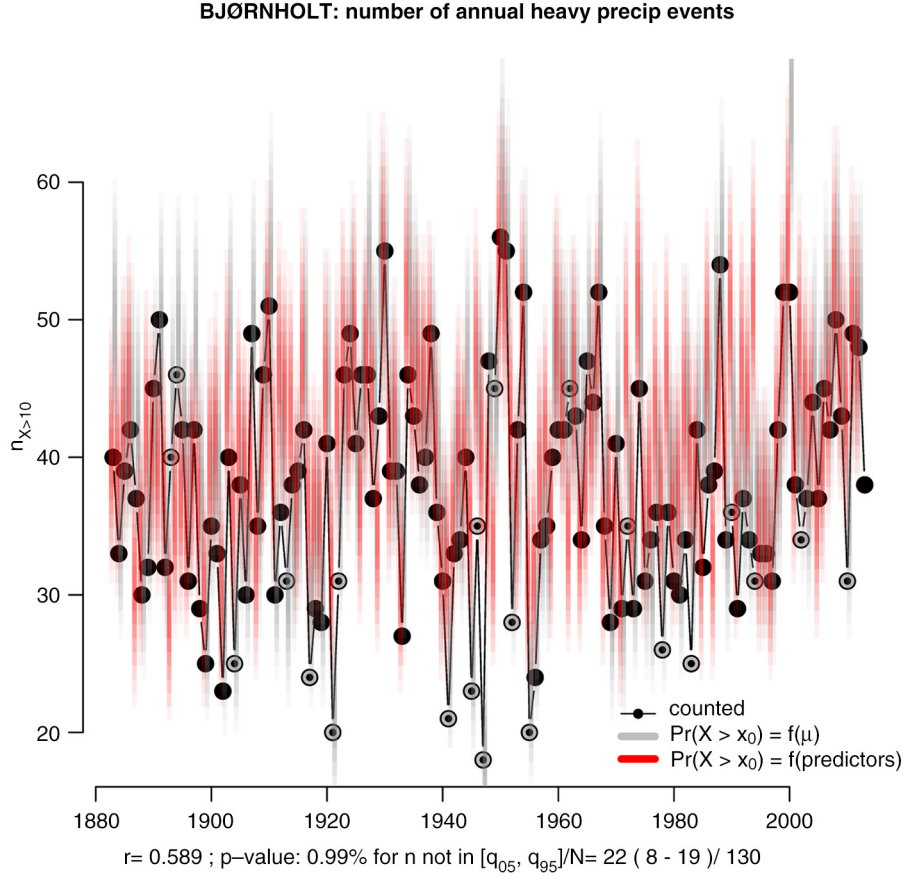


Fig. 9. Predicted and observed number of intense precipitation events, defined by exceeding the threshold value $x_0 = 10\text{mm/day}$. Symbols represent the counted events based on rain gauge data, grey shading the 90% confidence interval implied by μ and f_w derived from the observations, and red shading the 90% confidence interval derived through the downscaling exercise. The confidence intervals were derived taking the 5 and 95% percentiles from a binomial distribution where the probability for exceeding the 10 mm/day was taken from eq. (A4). The filled symbols mark the years when the number of events was outside the 90% confidence interval.

mean that the analysis presented here is not quite equal to ‘perfect predictor experiments’.

Different experiments can shed light on processes connecting the large-scale conditions and the local rainfall statistics, such as ‘pseudo-reality’-based ones. Such a study is beyond the scope of this paper, which addresses the question whether we can make actually useful forecasts of precipitation statistics seasons to decades ahead, given large-scale patterns in surface temperature and mean SLP. A ‘pseudo-reality’ set-up could also by-pass problems such as imperfections in the reanalysis data.

A concern for seasonal-to-decadal forecasting is whether GCMs are able to skilfully forecast the large-scale patterns in SAT and SLP over the North-Atlantic on a seasonal-to-decadal time scales, and the greatest bottle neck in terms of providing such forecasts for the North-Atlantic region seems to be due to the limited understanding of the role of various physical processes, the presence of complex chaos, and

whether the GCMs actually capture crucial precursory signals. The coarse spatial resolution in the ocean models used in the GCMs may limit their ability to adequately represent ocean currents and associated heat transport. Furthermore, Benestad et al. (2010) carried out a number of simulations with different Arctic sea-ice conditions and found that the temperature response was pronounced for Europe but not systematic, suggesting a strong presence of non-linearities. European teleconnection mechanisms may involve dynamical processes which are hard to forecast, such as the Arctic vortex, the NAO, jet streams, and storm tracks. The roles of oceanic heat distribution and currents are still not well known. In this study, the same predictor set-up was used for all the European stations, however, the skill will probably be higher for some locations with a different predictor choice. Seasonal forecasts tend to have more skill in the Tropics, e.g. for the ENSO, and the same methods used here may provide useful estimates of probabilities for heavy rainfall seasons ahead.

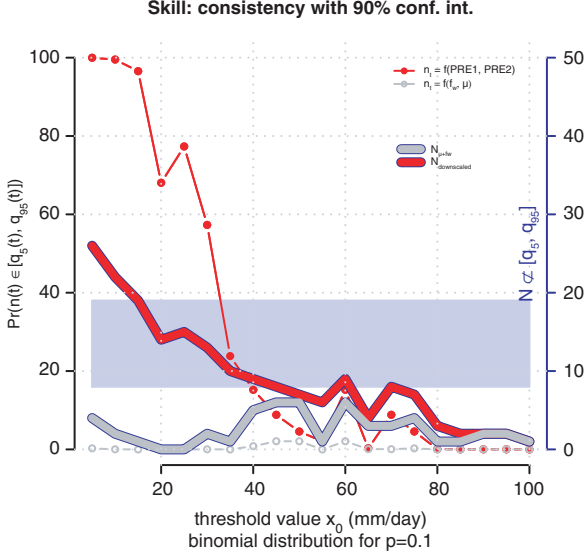


Fig. 10. A measure of skill. The right axis and the thick line indicate the number of observed events that were outside the predicted 90% confidence interval for a range of different threshold values x_0 (x-axis). The left axis and thin lines indicate the probability associated with the observed number of intense wet events outside the predicted 5–95% annual interval estimated according to a binomial distribution. The number of years with events outside the confidence interval are counted and then the probability is estimated according to the binomial distribution, the sample size (number of years), and a probability of 10% for being outside the predicted 90% confidence interval. The red curves show results from the downscaling exercise whereas the grey curves show the results derived from μ and f_w estimated directly from the rain gauge data.

5. Conclusions

We show how ESD can be applied to parameters that traditionally have not been commonly used in downscaling, such as predicting probabilities and number of intense precipitation events. Furthermore, we provide a contribution to a more ‘comprehensive evaluation’ (Murphy et al., 2010) to measure the compound effect of downscaling biases in terms of the number of events predicted over time. However, we do not attempt to assess the compound skill of GCM and downscaling here.

The results presented here suggest that it is possible to estimate local precipitation statistics with some predictive skill given large scale SAT and SLP conditions. The wet-day frequency in particular seems to bear a strong dependency to the large-scale atmospheric conditions, although a weak link was also found for the wet-day mean μ which is a proxy for the precipitation amount. A prediction of precipitation statistics on seasonal-to-decadal time scales based on these results will be probabilistic with approx-

imate estimates, and hence the presence of uncertainties will linger on. The precipitation amount (μ) was sensitive to time scale in certain circumstances and hence sample size. A possible reason for this may be the presence of different physical processes with different intrinsic time scales, most likely in the large-scale predictors and seasonal differences between processes producing rain and snow. Hence, the different time scales in the predictors and the predictands may reflect different physical processes in terms of the rainfall amount on a wet day. The weak skill for winter may also be due to the generally low precipitation amounts during winter and that the annual wet-day mean emphasises the season with higher rainfall. The wet-day frequency was generally not sensitive to seasonal or annual time scales, which may suggest that similar physical mechanisms responsible for the wet-dry occurrence is present on seasonal and annual time scales. The analysis for 5-yr time scales failed to find dependencies between large and small scales for the chosen predictors. The results suggest that models calibrated on annual data samples may in some cases provide skilful forecasts on seasonal time scales only.

6. Acknowledgements

This work has been supported by the Norwegian Meteorological Institute and SPECS (EU Grant Agreement 3038378). We also like to thank Reidun Gangstø, Inger Hanssen-Bauer and Oskar Andreas Landgren for reading and commenting on the manuscript.

7. Appendix

Definitions

The wet-day mean is the mean value of all days where the amounts exceed the threshold value $x_0 = 1\text{mm/day}$:

$$\mu = \sum_{i=1}^N (x_i) / n_w; \forall x > x_0, \quad (\text{A1})$$

where N is the rain gauge sample size (number of days) and the total number of wet days is $n_w = \sum_{i=1}^N \mathcal{H}(x_i > x_0)$ where \mathcal{H} is the Heaviside function. The wet-day frequency is the number of days where the amounts exceed the threshold value $x_0 = 1\text{mm/day}$ divided by the total number of days:

$$f_w = \sum_{i=1}^N \mathcal{H}(x_i > x_0) / N \quad (\text{A2})$$

Method in detail

Daily precipitation data were sorted into wet ($X > 1\text{mm/day}$) and dry days as in Benestad et al. (2012a, 2012b) and Benestad (2013), where the precipitation statistics were characterised by the annual wet-day mean μ and the annual

wet-day frequency f_w . The ordinary all-day mean precipitation is the product of the wet-day frequency and mean $\bar{x} = f_w \mu$.

The wet-day 24-hour precipitation was taken to be approximately exponentially distributed, whereas the variable describing a wet or a dry day was expected to follow a binomial distribution ($o \in [0,1]$). In contrary to the variables, statistical parameters describing the *annual mean* wet-day mean μ and the *annual mean* wet-day frequency f_w follow a normal distribution according to the central limit theorem (Wheelan, 2014), as seen in the qq-plots in Figs. 2 and 3. The design of the experiment where regression analysis was applied to annually aggregated statistics was motivated by the convenience of an expected normal distribution according to the central limit theorem.

The experiments used regression-based ESD, involving step-wise screening applied to the principal components describing the EOFs, and the regression model used in the ESD was

$$\mu(t) = \beta_0 + \sum_{i=1}^m (\beta_i x_{SAT}(t)_i) + \varepsilon, \quad (\text{A3})$$

where $x_{SAT}(t)$ refers to the principal components associated with the EOFs describing the annual mean SAT, and ε is an error term describing the residual. Equation A3 was repeated after μ was replaced by the residual ε and $x_{SAT}(t)$ with the principal components describing the SLP ($x_{SLP}(t)$) and the final downscaled results were the sum of the two steps.

On a separate account, we wanted to examine the possibility to predict reliable probabilities for extreme precipitation events, given a skilful prediction of μ and based on the prediction of pdfs for 24-hour precipitation. Let the 24-hour precipitation be approximately represented by an exponential distribution,

$$Pr(X < x) = \int_0^x \lambda e^{-\lambda x} dx = (1 - e^{-\lambda x}), \quad (\text{A4})$$

for which the mean value is $\mu = 1/\lambda$ and variance $\sigma^2 = 1/\lambda^2 = \mu^2$. If the wet-day precipitation follows an exponential distribution, then any quantile q_p is directly specified in terms of the wet-day mean μ :

$$q_p = -\ln(1 - p)\mu. \quad (\text{A5})$$

We also can solve for the value of any return value x_τ analytically if we know the fraction of wet days f_w and the data actually follows an exponential distribution (we know that this is not the case here, as the tail is thicker than for the exponential distribution):

$$\frac{1}{\tau} = f_w Pr(X > x_\tau) = f_w e^{-x_\tau/\mu}, \quad (\text{A6})$$

therefore,

$$x_\tau = -\ln\left(\frac{1}{\tau f_w}\right)\mu. \quad (\text{A7})$$

We can then estimate the probability of exceedance $X > x_\tau$ for a year and the number of events according to a binomial distribution. We let the probability for an event $X > x_\tau$ be $p = f_w e^{-x_\tau/\mu}$, and use the binomial distribution to estimate the number of events k estimated for a given interval with length n : $C_n^k p^k (1-p)^{(n-k)}$.

It is possible to compare x_τ from the equation above with the return values estimated from proper EVT using general extreme value distribution (GEV) (Coles, 1999) (Fig. A1). The return values for Bjørnholt was estimated using the R-package 'evd' and fgev which subsequently were compared with x_τ .

The R-package 'esd'

The wet-day mean μ and frequency f_w were derived from a set of methods provided by the R-package 'esd' (Benestad et al., 2014) developed under the R core programming tool (R Core Team, 2014):

```
mu <- as.annual(y, FUN = 'exceedance', threshold = threshold)
fw <- as.annual(y, FUN = 'wetfreq', threshold = threshold)
```

The 'esd' package is made freely available by the Norwegian Meteorological Institute (MET Norway) and is for use by

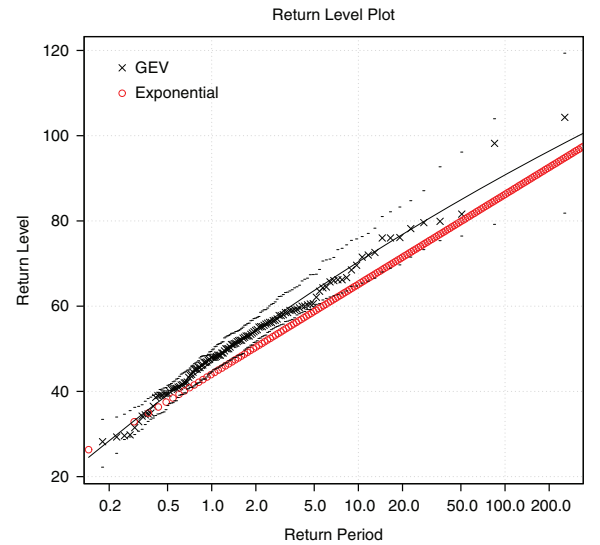


Fig. A1. A comparison between return values derived from GEV and eq. (A6) for Bjørnholt suggests that the exponential distribution has a tendency to under-estimate the return-levels. Still, the 'error' associated with the estimates based on the simpler expression for x_τ are less than 10% compared with the GEV-based estimates (not shown). Hence, a rough estimate for return-levels can be obtained by downscaling the wet-day mean μ and frequency f_w .

the climate community. It was primarily built for statistical downscaling of climate variables and parameters from global (e.g. ENSEMBLES, CMIP3/5 model results) and/or regional climate models (EURO-CORDEX), and has recently been extended to deal with data I/O, statistical analysis and visualisation. It consists of (1) retrieving and manipulating large samples of climate and model data from various sources, and (2) computing Empirical Orthogonal Functions (EOFs), PCA, and Multi-Variate Regressions (MVRs) using very simple R commands and procedures (See R Script in the supporting material). The esd package has been built with the emphasis of traceability, compatibility, and transparency of the data, methods, procedures, and results. As the ‘esd’ package has been built on R system, it inherits from the large number of R built-in functions and procedures. Finally, it can be easily tailored to various climate analysis with few adaptations such as the one described in this paper.

An example of programming commands is described hereafter including a four-steps procedure. The first step consists of retrieving daily precipitation from the MET Norway archive using the function ‘station’ as the following

```
y <- station(src = 'metnod', param = 'precip')
```

The second step consist of getting global climate model output and reanalysis data using the function ‘retrieve’ as the following

```
gcm <- retrieve('any global climate model'),
rea <- retrieve('any reanalysis').
```

The third step consists of computing the common EOFs using the function ‘EOF’ as follows

```
ceof <- EOF(combine(rea, gcm)).
```

Finally, the downscaling of y from the common EOF ‘ceof’ is achieved using the function ‘DS’ as the following

```
ds <- DS(y, ceof).
```

References

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**(6), 716–723. DOI: 10.1109/TAC.1974.1100705.
- Baldwin, M. P. and Dunkerton, T. J. 2001. Stratospheric harbingers of anomalous weather regimes. *Science*. **294**, 581–584.
- Baldwin, M. P., Stephenson, D. B., Thompson, D. W. J., Dunkerton, T. J., Charlton, A. J. and co-authors. 2003. Stratospheric memory and skill of extended-range weather forecasts. *Science*. **301**, 636–640.
- Barnett, T. and Preisendorfer, R. 1987. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.* **115**, 1825–1850.
- Barnett, T. P. 1999. Comparison of near-surface air temperature variability in 11 coupled global climate models. *J. Clim.* **12**, 511–518.
- Barnston, A. G., van den Dool, H. M., Zebiak, S. E., Barnett, T. P., Ji, M. and co-authors. 1994. Long-lead seasonal forecasts – where do we stand? *Bull. Am. Meteorol. Soc.* **75**, 2097–2114.
- Barsugli, J. J., Guentchev, G., Horton, R. M., Wood, A., Mearns, L. O. and co-authors. 2013. The practitioner’s dilemma: how to assess the credibility of downscaled climate projections. *Eos Trans. Am. Geophys. Union*. **94**(46), 424–425. DOI: 10.1002/2013EO460005.
- Benestad, R., Nychka, D. and Mearns, L. O. 2012a. Spatially and temporally consistent prediction of heavy precipitation from mean values. *Nat. Clim. Change*. **2**, 544–547. DOI: 10.1038/NCLIMATE1497.
- Benestad, R., Nychka, D. and Mearns, L. O. 2012b. Specification of wet-day daily rainfall quantiles from the mean value. *Tellus A*. **64**, 14981. DOI: 10.3402/tellusa.v64i0.14981.
- Benestad, R. E. 2001. A comparison between two empirical downscaling strategies. *Int. J. Climatol.* **21**, 1645–1668. DOI: 10.1002/joc.703.
- Benestad, R. E. 2013. Association between trends in daily rainfall percentiles and the global mean temperature. *J. Geophys. Res. Atmos.* **118**(19), 10802–10810. DOI: 10.1002/jgrd.50814.
- Benestad, R. E., Hanssen-Bauer, I. and Chen, D. 2008. *Empirical-Statistical Downscaling*. World Scientific, Singapore.
- Benestad, R. E. and Melsom, A. 2002. Is there a link between the unusually wet autumns in southeastern Norway and SST anomalies? *Clim. Res.* **23**, 67–79.
- Benestad, R. E., Mezghani, A. and Parding, K. M. 2014. esd for Mac & Linux/esd for windows. *figshare*. Online at: http://figshare.com/articles/esd_for_Mac_amp_Linux/1160493
- Benestad, R. E., Senan, R., Balmaseda, M., Ferranti, L., Orsolini, Y. and co-authors. 2010. Sensitivity of summer 2-m temperature to sea ice conditions. *Tellus A*. **63**, 324–337. DOI: 10.1111/j.1600-0870.2010.00488.x.
- Brönnimann, S., Luterbacher, J., Staehelin, J., Svendby, T. M., Hansen, G. and co-authors. 2004. Extreme climate of the global troposphere and stratosphere in the 1940–42 related el Niño. *Nature*. **431**, 971–974.
- Buishand, T. A. 1978. Some remarks on the use of daily rainfall models. *J. Hydrol.* **36**(3–4), 295–308. DOI: 10.1016/0022-1694(78)90150-6.
- Burgueo, A., Martnez, M. D., Lana, X. and Serra, C. 2005. Statistical distributions of the daily rainfall regime in Catalonia (northeastern Spain) for the years 1950–2000. *Int. J. Climatol.* **25**(10), 1381–1403. DOI: 10.1002/joc.1197.
- Cagnazzo, C. and Manzini, E. 2009. Impact of the stratosphere on the winter tropospheric teleconnections between ENSO and the north Atlantic and European region. *J. Clim.* **22**, 1223–1238. DOI: 10.1175/2008JCLI2549.1.
- Coles, S. 1999. Extreme value theory and applications. In: *Notes from a Course on EVT and Applications Presented at the 44th Reunio Annual da RBRAS e 8th SEAGRO*, Bucato, So Paulo, Brazil, 26–30 July, 1999. Online at: <http://www.maths.bris.ac.uk/masgc/extremes/exnewnotes2.ps>

- Compo, G. P. and Sardeshmukh, P. D. 2004. Storm track predictability on seasonal and decadal scales. *J. Clim.* **17**, 3701–3720.
- Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J. and co-authors. 2011. The twentieth century re-analysis project. *Q. J. Roy. Meteorol. Soc.* **137**(654), 1–28. DOI: 10.1002/qj.776.
- Derome, J., Brunet, G., Plante, A., Gagnon, N., Boer, G. J. and co-authors. 2001. Seasonal predictions based on two dynamical models. *Atmos. Ocean.* **39**(4), 485–501.
- Derome, J., Lin, H. and Brunet, G. 2005. Seasonal forecasting with a simple general circulation model: predictive skill in the AO and PNA. *J. Clim.* **18**, 597–611.
- Dix, M. R. and Hunt, B. G. 1995. Chaotic influences and the problem of deterministic seasonal predictions. *Int. J. Climatol.* **15**(7), 729–752. DOI: 10.1002/joc.3370150703.
- Doblas-Reyes, F. J., Hagedorn, R. and Palmer, T. N. 2007. Developments in dynamical seasonal forecasting relevant to agricultural management. *Clim. Res.* **33**, 19–26.
- Fedderson, H. 2003. Predictability of seasonal precipitation in the Nordic region. *Tellus A.* **55**, 385–400.
- Fedderson, H. and Andersen, U. 2005. A method for statistical downscaling of seasonal ensemble predictions. *Tellus A.* **57**, 398–408.
- Frauenfelder, R., Solheim, A., Isaksen, K., Romstad, B., Dyrørdal, A. V. and co-authors. 2013. *Impacts of Extreme Weather Events on Infrastructure in Norway (InfraRisk)*. Sluttrapport til NFR – prosjekt 200689 NGI rapport nr. 20091808. Norges Geotekniske Institutt, Oslo, Norway.
- George, S. E. and Sutton, R. T. 2006. Predictability and skill of boreal winter forecasts made with the ECMWF seasonal forecast system II. *Q. J. Roy. Meteorol. Soc.* **132**, 2031–2053.
- Gershunov, A. and Barnett, T. P. 1998. ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: observations and model results. *J. Clim.* **11**, 1575–1586.
- Hanssen-Bauer, I., Frland, E. J., Haugen, J. E. and Tveito, O. E. 2003. Temperature and precipitation scenarios for Norway: comparison of results from dynamical and empirical downscaling. *Clim. Res.* **25**, 15–27.
- Hawkins, T. W., Ellis, A. W., Skindlov, J. A. and Reigle, D. 2002. Inter-annual analysis of the North American snow cover-monsoon teleconnection: seasonal forecasting utility. *J. Clim.* **15**, 1743–1753.
- Johansson, Å. 2007. Prediction skill of the NAO and PNA from daily to seasonal time scales. *J. Clim.* **20**, 1957–1975.
- Johansson, Å., Barnston, A., Saha, S. and van den Dool, H. 1998. On the level and origin of seasonal forecast skill in northern Europe. *J. Atmos. Sci.* **55**, 103–127.
- Kharin, V. V. and Zwiers, F. W. 2003. Improved seasonal probability forecasts. *J. Clim.* **16**, 1684–1701.
- Klein Tank, A. J. B. W., Können, G. P., Böhm, R., Demarée, G., Gocheva, A. and co-authors. 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. *Int. J. Climatol.* **22**, 1441–1453.
- Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z. and co-authors. 1999. Improved weather and seasonal climate forecasts from multimodel superensembles. *Science*. **285**, 1548–1550.
- Landman, W. A., Kgatuke, M.-J., Mbedzi, M., Beraki, A., Bartman, A. and co-authors. 2009. Performance comparison of some dynamical and empirical downscaling methods for South Africa from a seasonal climate modelling perspective. *Int. J. Climatol.* **29**, 1535–1549. DOI: 10.1002/joc.1766.
- Li, S., Goddard, L. and DeWitt, D. G. 2008. Predictive skill of AGCM seasonal forecasts subject to different SST prediction methodologies. *J. Clim.* **21**, 2169–2186.
- Linderson, M.-L., Achberger, C. and Chen, D. 2004. Statistical downscaling and scenario construction of precipitation in Scania, southern Sweden. *Nord. Hydrol.* **35**, 261–278.
- Lloyd-Hughes, B. and Saunders, M. A. 2002. Seasonal prediction of European spring precipitation from El Niño-southern oscillation and local sea-surface temperatures. *Int. J. Climatol.* **22**, 1–14.
- Lorenz, E. N. 1956. *Empirical Orthogonal Functions and Statistical Weather Prediction*. Scientific Report No. 1, Department of Meteorology, MIT, Cambridge, MA.
- Maraun, D., Widmann, M., Gutierrez, J., Kotlarski, S., Chandler, R. and co-authors. 2015. Value – a framework to validate downscaling approaches for climate change studies. *Earths Future*. **3**, 1–14. DOI: 10.1002/2014EF000259.
- Murphy, J., Kattsov, V., Keenlyside, N., Kimoto, M., Meehl, G. and co-authors. 2010. Towards prediction of decadal climate variability and change. *Proc. Environ. Sci.* **1**, 287–304. DOI: 10.1016/j.proenv.2010.09.018.
- North, G. R., Bell, T. L. and Cahalan, R. F. 1982. Sampling errors in the estimation of empirical orthogonal functions. *Mon. Weather Rev.* **110**, 699–706.
- Orsolini, Y., Senan, R., Benestad, R. E. and Melsom, A. 2011. Autumn atmospheric response to the 2007 low Arctic sea ice extent in coupled ocean-atmosphere hindcasts. *Clim. Dyn.* **38**, 2437–2448.
- Palmer, T. 2004. Progress towards reliable and useful seasonal and interannual climate prediction. *WMO Bull.* **53**, 325–333.
- Palmer, T., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.
- Palmer, T. and Anderson, D. 1994. The prospect for seasonal forecasting – a review paper. *Q. J. Roy. Meteorol. Soc.* **120**, 755–793.
- Palmer, T. N., Brankovic, C. and Richardson, D. S. 2000. A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. Roy. Meteorol. Soc.* **126**, 2013–2034.
- R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Core Team R Foundation for Statistical Computing, Vienna, Austria.
- Reason, C. J. C., Landman, W. and Tennant, W. 2006. Seasonal to decadal prediction of southern African climate and its links with variability of the Atlantic ocean. *Bull. Am. Meteorol. Soc.* **87**, 941–955.
- Rodwell, M. J. and Doblas-Reyes, F. J. 2006. Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Clim.* **19**, 6025–6045.

- Stockdale, T. N., Anderson, D. L. T., Alves, J. O. S. and Balmaseda, M. A. 1998. Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*. **392**, 370–373.
- Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K. and co-authors. 2013. Climate change 2013: the physical science basis. In: *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK.
- Strang, G. 1988. *Linear Algebra and its Application*. Harcourt Brace & Company, San Diego, CA.
- Thom, H. C. S. 1958. A note on the gamma distribution. *Mon. Weather Rev.* **86**(4), 117–122.
- Thompson, D. W. J., Baldwin, M. P. and Wallace, J. M. 2002. Stratospheric connection to northern Hemisphere wintertime weather: implications for prediction. *J. Clim.* **15**, 1421–1428.
- Ueda, H., Shinoda, M. and Kamahori, H. 2003. Spring northward retreat of Eurasian snow cover relevant to seasonal and inter-annual variations of atmospheric circulation. *Int. J. Climatol.* **23**, 615–629.
- van der Linden, P. and Mitchell, J. F. B. (eds.). 2009. *Ensembles: Climate Change and its Impacts: Summary of Research and Results from the ENSEMBLES Project European Commission*. Met Office Hadley Centre, Exeter, UK.
- Vitard, F., Stockdale, T. and Ferranti, L. 2007. Seasonal forecasting of tropical storm frequency. *ECMWF Newsllett.* **112**, 16–22.
- Weigel, A. P., Liniger, M. A. and Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. Roy. Meteorol. Soc.* **134**(630), 241–260. DOI: 10.1002/qj.210.
- Wheeler, C. J. 2014. *Naked Statistics: Stripping the Dread from the Data*. Met W. W. Norton & Company, New York, NY.
- Wijngaard, J. B., Tank, A. M. G. K. and Können, G. P. 2003. Homogeneity of 20th century daily temperature and precipitation series. *Int. J. Climatol.* **23**, 679–692.
- Wilby, R. and Wigley, T. 2000. Precipitation predictors for downscaling: observed and general circulation model relationships. *Int. J. Climatol.* **20**, 641–661.
- Wilby, R. L. and Wigley, T. M. L. 2002. Future changes in the distribution of daily precipitation totals across north America. *Geophys. Res. Lett.* **29**(7), 39-1–39-4. DOI: 10.1029/2001GL013048.
- Wilks, D. 1998. Multisite generalization of a daily stochastic precipitation model. *J. Hydrol.* **210**, 178–191.
- Wilks, D. S. 1995. *Statistical Methods in the Atmospheric Sciences*. Academic Press, Orlando, FL.
- Wilks, D. S. 2006. On “field significance” and the false discovery rate. *J. Appl. Meteorol. Climatol.* **45**, 1181–1189. DOI: 10.1175/JAM2404.1.
- Wilks, D. S. and Wilby, R. L. 1999. The weather generation game: a review of stochastic weather models. *Prog. Phys. Geogr.* **23**(3), 329–357. DOI: 10.1177/030913339902300302.
- Wolff, M., Isaksen, K., Petersen-Øverleir, A., Ødemark, K., Reitan, T. and co-authors. 2014. Derivation of a new continuous adjustment function for correcting wind-induced loss of solid precipitation: results of a Norwegian field study. *Hydrol. Earth Syst. Sci. Discuss.* **11**, 10043–10084. DOI: 10.5194/hessd-11-10043-2014.
- Yun, W. T., Stefanova, L. and Krishnamurti, T. N. 2003. Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Clim.* **16**, 3834–3840.
- Zhang, H. and Frederiksen, C. S. 2003. Local and nonlocal impacts of soil moisture initialization on AGCM seasonal forecasts: a model sensitivity study. *J. Clim.* **16**, 2117–2137.