Spatially and temporally consistent prediction of heavy precipitation from mean values

R. E. Benestad[1]*, D. Nychka[2] and L. O. Mearns[2]

Extreme precipitation can cause flooding, result in substantial damages and have detrimental effects on ecosystems[1,2]. Climate adaptation must therefore account for the greatest precipitation amounts that may be expected over a certain time span[3]. The recurrence of extreme-to-heavy precipitation is notoriously hard to predict, yet cost–benefit estimates of mitigation and successful climate adaptation will need reliable information about percentiles for daily precipitation. Here we present a new and simple formula that relates wet-day mean precipitation to heavy precipitation, providing a method for predicting and downscaling daily precipitation statistics. We examined 32,857 daily rain-gauge records from around the world and the evaluation of the method demonstrated that wet-day precipitation percentiles can be predicted with high accuracy. Evaluations against independent data demonstrated high skill in both space and time, indicating a highly robust methodology.

Traditionally, design values, used for dimensioning infrastructure or specifying insurance premiums, have been estimated from statistics such as extreme-value theory or percentiles. Such statistical analyses usually assume that the statistical distribution is constant, but this assumption breaks down if there is a climate change where the statistical character of rainfall is non-stationary. Some extreme-value methods, however, can also account for non-stationary statistics, but the parameters used to fit the extreme-value distributions (scale, shape and location) often do not have a clear connection to physical processes. Such methods are also based on a number of subjective choices, such as between peak-over-threshold or block maxima[4], and are difficult to incorporate into a downscaling framework accounting for climatic changes.

Previous downscaling efforts for extreme events have often focused on certain extreme indices[5,6], extreme-value distributions[7,8], or involved attempts to directly downscale the entire 24-h precipitation probability distribution[9]. Regional climate models (RCMs) have also been used to predict heavy precipitation and changes in such precipitation[7], however, they do not provide the same kind of description of the precipitation as observed as they describe an area mean as opposed to point measurements. It is also important to keep in mind that models in principle are idealistic simplifications and are designed to describe the most essential features of the climate, not all the small details. The description of cloud microphysics in RCMs is still limited in terms of representing small-scale processes that are suspected to take place[10]. Evaluations of some RCMs have also indicated that they are not able to provide a realistic description of the higher quantiles[11] and that they produce drizzle conditions too frequently[12]. Hence, adequate modelling of precipitation statistics has been a long-standing problem.
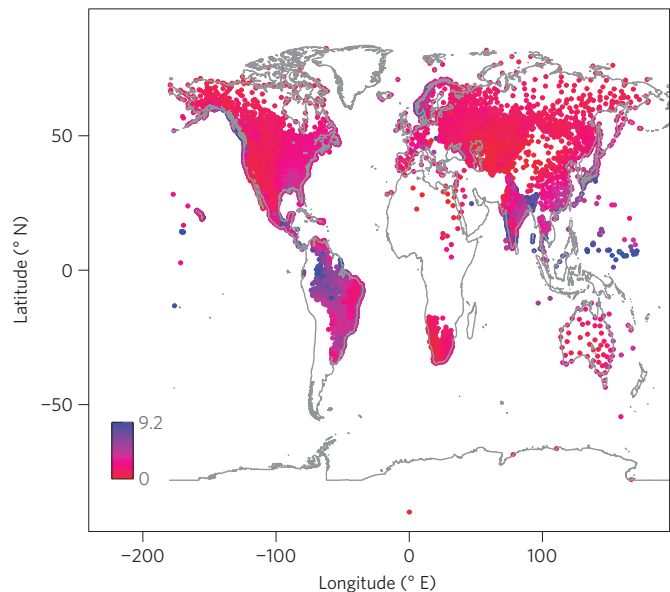

Figure 1 | Map showing the locations of the 33,599 original rain gauges on which the analysis was based. The results of the study are representative for a large part of the world, but there are also some regions for which there were no available data (the world oceans, Africa, the Middle East, Oceania and part of Latin America). The colour coding shows the mean precipitation (mm d^{-1}) for the season June–September.

Recently, there have been some new approaches for describing extreme precipitation, some of which have involved principal component analysis (PCA)[13]. Other efforts have tried to make use of the similarity between precipitation and exponential distribution for wet days to provide a crude approximation of the upper percentiles[9,14]. The advantage of these new approaches is that they can be more easily used in downscaling of global climate models. We extended these earlier efforts by combining them to provide an accurate description of the wet-day percentiles. The PCA can then be used together with regression analyses to predict percentiles for the daily wet-day precipitation over most of the Earth. This strategy transforms the data to reveal a robust and simple formula that directly links the wet-day percentiles to the wet-day mean precipitation (see Methods).

We examined 32,857 time series of daily precipitation from all over the world, taken from the Global Daily Climatology Network[15,16] and European Climate Assessment & Dataset project[17] data sets, as well as the Norwegian Meteorological Institute's climate

[1]The Norwegian Meteorological Institute, Oslo 0313, Norway, [2]National Center for Atmospheric Research, Boulder, Colorado 80305, USA.
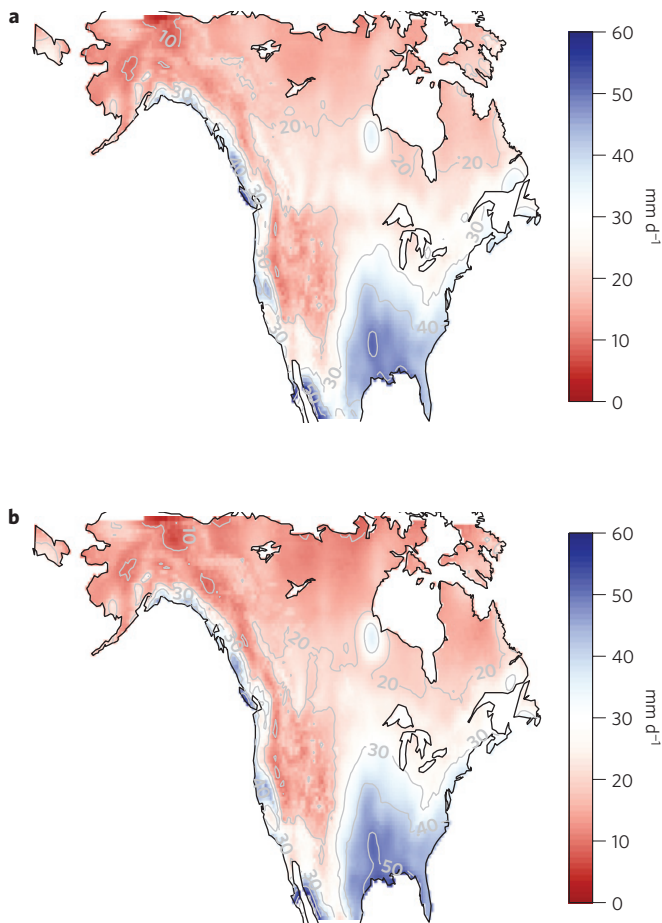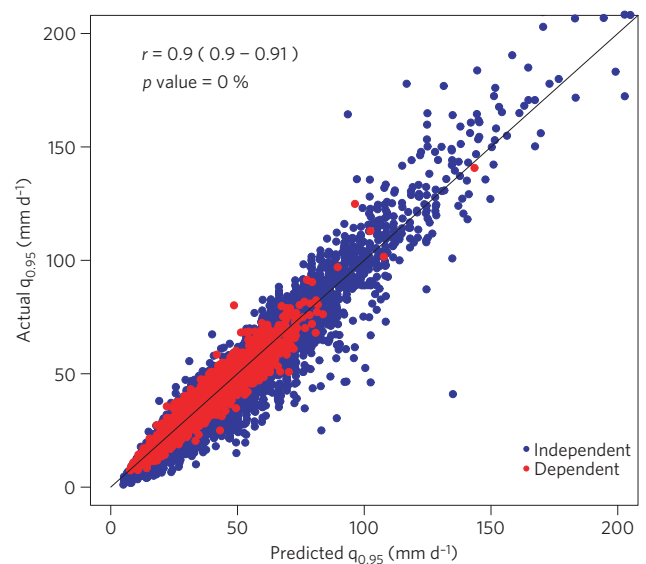*e-mail: rasmus.benestad@met.no.

Figure 3 | Scatter plot providing an evaluation of the predicted wet-day $q_{0.95}$ values for dependent data from North America presented in Fig. 2 (red) and independent data representing rain gauges outside North America (blue). The predictions are shown along the *x* axis and the observations along the *y* axis.

Figure 2 | Maps of wet-day 95th percentile for 24-h precipitation a, The map shows the raw wet-day $q_{0.95}$ as extracted from the Global Daily Climatology Network rain-gauge data. Here $q_{0.95}$ was estimated for each respective location, based on the months of June–September for the entire record. The map was then generated by gridding $q_{0.95}$ for each location, using the R-package LatticeKrig (see Supplementary Information) and *z* as a covariate. The grid resolution is 0.65° longitude by 0.45° latitude. **b**, As in **a**, but showing predicted wet-day $q_{0.95}$ values based on μ, f_w, *z* and *d*. These predictions involved both a regression analysis and a PCA applied to a range of quantiles as explained in the text and ref. 13. The correlation between predicted and observed values are presented in Fig. 3 as a scatter plot of red symbols.

archive[18]. The location of the rain gauges are shown in Fig. 1; the colour coding refers to each location's mean precipitation. Stations with fewer than 1,000 wet-day amounts were discarded, keeping 14,022 stations from North America and 14,003 stations from the rest of the world. Here the threshold for defining a wet day was taken to amounts exceeding 1 mm d^{-1} and the mean wet-day value is represented by μ.

Percentiles are here represented by q_p and are used to describe the value that is greater than the *p* portion of the sampled data. They also describe the amount that can be associated with a probability: $P(X > q_p) = 1 - p$. We will distinguish between the statistics estimated only for the days with precipitation (wet days) and the statistics describing both dry and wet days.

For downscaling purposes, we wanted to explore the relationship between the two leading principal components and local conditions. We expected that μ may be one important factor, as any percentile in an exponential distribution is entirely specified by μ. We also suspected that the wet-day frequency (f_w), distance to the coast (*d*) and elevation (*z*) may be factors that affect the

precipitation characteristics. An ordinary linear regression was carried out on the two leading principal components describing the precipitation data from North America. Predictions of $q_{0.95}$ were then made by first multiplying the regression coefficients with μ, f_w, *d*, and *z* to estimate the two leading principal components, and then using these principal components together with the empirical orthogonal functions (EOFs) and eigenvalues to reconstruct the percentiles. The predictions based on μ, f_w, *d* and *z* are shown in Fig. 2b, which reproduce most of the features seen in the original observations (Fig. 2a; the correlation is presented as the red symbols in Fig. 3). For this comparison, the model is evaluated on the same data set as that used to train the method; in other words, a dependent data set.

To test how well this regression model was able to predict values for out-of-sample data, we repeated the PCA, but now applied to the data set for the whole world (the reason for this is explained in the Supplementary Information). The part of the principal components that represented North American locations were subsequently extracted and used to calibrate the regression model (referred to as dependent data). The remaining part of the principal components was not used in the regression analysis, but as independent data for evaluation purposes. Figure 3 shows a scatter plot for the predicted values of $q_{0.95}$ versus the observed values both for the set of dependent (red; also shown in Fig. 2) and independent data (blue). The correlation between the predicted and observed independent data was 0.93 and the likelihood that this result should arise from chance was virtually zero (the size of the data set, $N = 14,003$). The evaluation against the independent data was extended by dividing the rain-gauge data from North America into two equally sized batches, where the PCA and model calibration were carried out on data from one, and the evaluation was based on data from the other (5,640 stations). The correlation between predicted and observed values for $q_{0.95}$ from this out-of-batch evaluation was 0.95. More details and maps of the observed and predicted values of $q_{0.95}$ for regions outside North America are provided in the Supplementary Information.

If this relationship is to be used for downscaling percentiles in a climate-change scenario, it is also important to assess whether
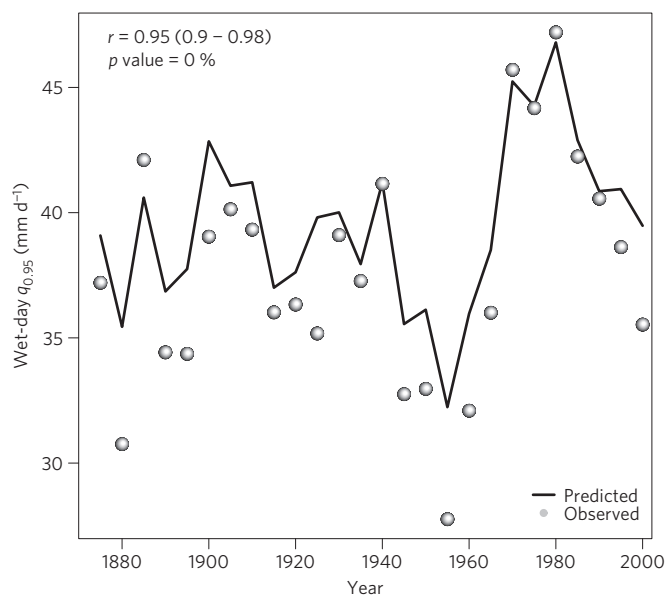
Figure 4 | A comparison between predicted (line) and observed (symbols) wet-day $q_{0.95}$ time series for the longest record of 24-h precipitation in the United States (station number 305801). The correlation between observed and predicted values is 0.95, with a 5–95% confidence range of 0.90–0.98. The regression model is the same as presented in Fig. 3 (red symbols) and was calibrated with 14,022 different rain-gauge records from North America. These results show that models calibrated with spatial variability also produce skillful predictions for the time dimension.

the relationship between the predictor and the predictand holds in time as well as in space. We picked the longest precipitation record from the United States and divided it into five-year sequences. For each sequence, we estimated μ, f_w and $q_{0.95}$. The values for d and z were constant and determined from the station's metadata. We multiplied these four parameters with the regression coefficients derived from the geographic distribution in μ, f_w, d and z from 14,003 North American rain gauges to predict local time-dependent values for the two leading principal components. The predicted principal component values were then used to reconstruct a time series of $q_{0.95}$. The actual percentiles are shown as symbols in Fig. 4, the predictions are shown as a black curve and the correlation between the two is 0.95. As the percentiles are only sampled from relatively short five-year intervals, some discrepancies may be expected owing to sampling fluctuations.

To see if the case presented in Fig. 4 was a special one, we repeated the procedure for all the North American rain data (11,281 stations), and compared the distribution of the correlation coefficients with similar analysis applied to random series. The mean correlation for the North American stations was 0.8, whereas for the random series it was 0.0 (see the Supplementary Information). We also carried out some simple sensitivity tests of our model and the results indicate that the percentiles are most sensitive to variations in μ, whereas the other three covariates induce only a secondary effect (Supplementary Information, Supplementary Fig. S10). The leading EOF has a shape that closely follows the diagonal in a quantile–quantile plot with the different axes representing $q_{0.95}$ and $q'_{0.95}$ respectively[14]; any scaling of the leading EOF, such as different leading principal component values, will have an effect of sliding the curve up or down along the diagonal. As the x axis represents $q'_p = -\ln(1-p)\mu$, μ must be closely linked to the leading principal component.

We have presented evidence for a close relationship between μ and q_p, which becomes apparent only when excluding non-rainy days, using the exponential distribution as a guide, and

transforming the data through a PCA. This relationship is valid in both space and time, and can be used to facilitate downscaling of daily precipitation statistics. The Supplementary Information provides a crude demonstration of how the exponential distribution, PCA, and μ can be used for downscaling (Supplementary Fig. S13). The results here also show that the key variables to simulate well for hydrological studies are μ and f_w. The projection of future precipitation statistics will require proper treatment of multimodel ensembles and Bayesian statistics, and is beyond the scope of this paper. Nevertheless, most global climate models predict wetter conditions over the mid-latitudes and parts of the tropics[19,20], and if this increase is reflected in the μ, as in the example in the Supplementary Information, then we can expect to see more extreme precipitation amounts in the future.

One question that has not been addressed is which physical processes may be linking μ to q_p and result in a distribution resembling an exponential with an increasing bias for higher percentiles. The rainfall statistics are a product of complex processes taking place on small spatial scales, involving rising air parcels, turbulent motion, entrainment, humidity, cold and warm initiation, and cloud-particle growth through condensation as well as collision and coalescence[21]. Furthermore, the precipitation amount at a given point depends on both the spatial extent and the speed of the system passing through. An interesting question, albeit beyond the scope of this paper, is whether the link between μ and q_p can be explained in terms of cloud environment, humidity, or microphysics, which most likely involves stochastic behaviour and cascading avalanches.

The link between μ and q_p also provides a number of useful applications for climate research. For instance, the formula can facilitate an effective compression of statistical information, as μ and the PCA results provide a complete description of the percentiles. Another aspect is that the formula can serve as a quality control to flag suspect data as most stations exhibit similar structures. It is also possible that it may benefit attribution analyses, when extreme rainfall may be associated with increases in μ. Owing to sampling fluctuations, it is easier to find long-term trends in μ than in higher percentiles, but given a direct link between the two, we can deduce an increased probability of extreme rainfall with increasing values of μ.

Methods

For each station, we extracted μ and 14 wet-day q_p values (estimated from the entire data record and taking p to be in (0.50, 0.99)), as well as keeping track of the number of wet and dry days, the location coordinates and z. The q_p values were combined into a vector with corresponding percentiles assuming an exponential distribution with the same value of μ. The latter are represented here by the symbol $q'_p = -\ln(1-p)\mu$ to distinguish them from the actual percentiles q_p. The relationship between the two types of percentiles is close to being one-to-one and a structure that follows the diagonal, albeit with a positive bias at higher values that becomes apparent when plotted in a scatter plot with q_p along the y axis and the corresponding q'_p on the x axis (see Supplementary Fig. S1 and ref. 14).

Then the 14 percentiles from all rain-gauge locations, stored in one vector for each site containing both q_p and q'_p, were combined into a data matrix for which the vectors made up the columns. This data matrix was subsequently subject to a PCA as in ref. 14. Here the PCA was implemented using a singular value decomposition that transformed the original data matrix into three components: EOFs, principal components and eigenvalues. The principal components represent the weight at each location that must be applied to the EOFs and eigenvalues to reconstruct any of the 14 quantiles. A detailed description of the PCA is provided in the Supplementary Information.

One advantage of this technique is that it provides a transform of the data by adding a perturbation to the exponential distribution. The leading EOF was by far the most important contribution (99% of variance) over North America, the second mode accounted for ~0.5% and the higher-order ones can be considered as noise. Hence we retained only the two leading EOFs for reconstructing the percentiles.

A regression analysis was carried out to predict the values for the two leading principal components based on μ, f_w, d and z. The predicted principal component values were then used together with the EOFs and the eigenvalues to derive the quantiles. Although the PCA spanned a range of quantiles between $q_{0.50}$ and

$q_{0.99}$, the present analysis highlights $q_{0.95}$ for precipitation because this quantity is commonly discussed in the literature on precipitation and $q_{0.99}$ implies greater random-sampling fluctuations and more noisy results[14]. A state-of-the art tool, using a kriging algorithm that takes the altitude as one covariate, was used to grid the observed wet-day 95th percentile $q_{0.95}$ for precipitation over North America in Fig. 2a (see Supplementary Information).

References

1. Min, S-K. *et al.* Human contribution to more-intense precipitation extremes. *Nature* **470,** 378–381 (2011).
2. Pall, P. *et al.* Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. *Nature* **470,** 382–385 (2011).
3. www.mdt.mt.gov/publications/docs/manuals/hyd/ch7.pdf.
4. Coles, S. G. *An Introduction to Statistical Modeling of Extreme Values* (Springer, 2001).
5. IPCC *Climate Change 2007: Impacts, Adaptation and Vulnerability* (eds Parry, M. L., Canziani, O. F., Palutikof, J. P., van der Linden, P. J. & Hanson, C. E.) (Cambridge Univ. Press, 2007).
6. Frich, P. *et al.* Observed coherent changes in climatic extremes during the second half of the twentieth century. *Clim. Res.* **19,** 193–212 (2002).
7. Frei, C. *et al.* Future change of precipitation extremes in Europe: Intercomparison of scenarios from regional climate models. *J. Geophys. Res.* **111,** D06105 (2006).
8. Wang, J. & Zhang, X. Downscaling and projection of winter extreme daily precipitation over North America. *J. Clim.* **21,** 923–937 (2008).
9. Benestad, R. E. Novel methods for inferring future changes in extreme rainfall over northern Europe. *Clim. Res.* **34,** 195–210 (2007).
10. O'Dowd, C. D. *et al.* The relative importance of non-sea-salt sulphate and sea-salt aerosol to the marine cloud condensation nuclei population: An improved multi-component aerosol–cloud droplet parametrization. *Quart. J. R. Meteorol. Soc.* **125,** 1295–1313 (1999).
11. Orskaug, E. *et al.* Evaluation of a dynamic downscaling of precipitation over the Norwegian mainland. *Tellus A* **63,** 746–756 (2011).
12. The BACC Author Team *Assessment of Climate Change for the Baltic Sea Basin* (Springer, 2008).
13. http://www.image.ucar.edu/~nychka/manuscripts/nychka_extremesEPFL.pdf.
14. Benestad, R. E., Nychka, D. & Mearns, L. O. Specification of wet-day daily rainfall quantiles from the mean value. *Tellus A* **64,** 14981 (2012).
15. Peterson, T., Daan, H. & Jones, P. Initial selection of a GCOS surface network. *Bull. Am. Meteorol. Soc.* **78,** 2145–2152 (1997).
16. http://www.ncdc.noaa.gov/oa/climate/research/gdcn/gdcn.html.
17. Klein Tank, A. *et al.* Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *Int. J. Climatol.* **22,** 1441–1453 (2002).
18. http://eklima.met.no.
19. Meehl, G. A. *et al.* in *IPCC Climate Change 2007: The Physical Science Basis* (eds Solomon, S. *et al.*) (Cambridge Univ. Press, 2007).
20. Benestad, R. E. Can we expect more extreme precipitation on the monthly time scale? *J. Clim.* **19,** 630–637 (2006).
21. Blyth, A. M. *et al.* Observation of supercooled raindrops in New Mexico summertime clouds. *J. Atmos. Sci.* **54,** 569–575 (1997).

Author contributions

R.E.B. contributed to project planning, implemented the R-package, carried out the experimental work and the data analysis and took charge in the writing-up process; D.N. took part in project planning, interpretation of results and provided the statistical and gridding expertise; L.O.M. participated in the project planing and discussion of results.

Additional information