International Journal of
Climatology

Global evaluation of surface temperature trends simulated by multi-model ensembles

Journal:	*International Journal of Climatology*
Manuscript ID:	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Benestad, Rasmus; The Norwegian Meteorological Institute, Climate Division
Keywords:	climate models, downscaling, evaluation, trends, ensembles

Benestad (2012), **International Journal of Climatology**, pp. 1–43

Global evaluation of surface temperature trends simulated by multi-model ensembles

By R.E. Benestad

The Norwegian Meteorological Institute, PO Box 43, 0313, Oslo, Norway *

First submission: June 19, 2012

Revised: .

* *Corresponding author: R.E. Benestad, rasmus.benestad@met.no, The Norwegian Meteorological Institute, PO Box 43, 0313 Oslo, Norway, phone +47-22 96 31 70, fax +47-22 96 30 50*

ABSTRACT

An ensemble of global climate models is evaluated for its skill in reproducing the past surface temperature trends seen in the observational data record. The evaluation is first carried out on a global scale and is then repeated for smaller spatial scales in order to shed light on the minimum spatial skillful scale. The world-wide set of downscaled results enable new ways of evaluating climate model, and the question of whether the ensemble is able to provide a skillful probabilistic description of the temperature trend is addressed. The results suggest that the global climate models provide skillful reproductions of the global mean warming over the past 50 years, but they tend to under-estimate the warming when the regions are approaching the size of $\sim 20 \times 10$ grid-boxes. The reduced ability to reproduce observed warming trends may also influence downscaled results, and an evaluation of empirical-statistical downscaled 2-m temperature world-wide suggests that the modelled values are roughly within the range of observed values. However, the statistical distributions are not identical, and the models appear to lose the ability to discern geographical differences in the trends as the spatial scales are reduced: They don't predict the cooling seen over the last decades at some locations. Furthermore, downscaled results tend to under-estimate the magnitude of the most pronounced trends.

KEY WORDS: Empirical-Statistical downscaling temperature trend evaluation multi-model ensembles minimum skillful scale

Introduction

The question of evaluating and estimating the skill associated with regional climate model results is a major topic in the COST ES1102 "VALUE" project*. Hulme and Dessai (2008) regarded the *process* of constructing scenarios as potentially beneficial irrespective of the value of end product, and pointed out that this framing yields different expectations about how one might evaluate the success of scenario exercises. They presented three approaches to evaluating the success or otherwise, which in practice can be considered at three different levels: predictive success, decision success and learning success. Nevertheless, the question about predictive success is an important aspect, and it is important to evaluate the skill of the multi-model ensembles downscaling exercises as one over-arching tool for predicting changes, rather than focusing on single models. Recent papers, such as Weigel *et al.* (2008) have pointed out some principal differences between single models and ensembles. It is argued that that multi-model combination may result in "overconfidence" (i.e. ensemble spread is widened while average ensemble-mean error is reduced) and the inclusion of even poor models may give an impression of increased skill.

Climate models used to guide climate change adaptation must demonstrate skill at the relevant spatial and temporal scales. A number of recent papers have presented critical arguments in terms of downscaling climate models in order to provide a detailed description of local climate necessary for climate change adaptation (Oreskes *et al.*, 2010; Pielke Sr. and Wilby, 2012). Palmer (2011) also argued that there still is insufficient computing power to solve the equations of climate science with sufficient accuracy, which requires a grid-point spacing of about a few tens of kilometres or less. However, the arguments presented in these papers were based on general principles and judgements,

* http://www.value-cost.eu/

and not specific evaluations of the models and the downscaling climate model results.

In order to make any statement about the models' ability to predict real-world features it is important to consider the exact definition of 'skill', and clarify what is implied in a poor and skillful model. There is a body of literature on skill estimation for forecasts of model ensembles, such as the Brier skill score (BSS), Relative operating characteristic (ROC), and reliability diagrams (Wilks, 1995; Jolliffe and Stephenson, 2003).

An evaluation of model results should reflect the quantity that is of interest in terms of prediction. When using climate models to make climate change scenarios for climate change adaptation, it is the ability to predict the local warming that is one of the most relevant issues, and the evaluation should include the whole the process of downscaling a multi-model ensemble and trend analysis. Such an evaluation needs to address the question whether the analysis of the an ensemble of downscaled global climate models is able to reproduce the historical trends seen in the observations (Benestad *et al.*, 2007). Furthermore, if climate models are used to answer questions about future heatwaves and cold snaps, the relevant question is whether the models have been able to describe a typical cold or warm season in the past. Before such questions can be addressed, it is important to examine whether observed low and high quantiles are reproduced by the climate models at the end of entire analytical process. It is important to keep in mind that the success of such an assessment is merely a minimum and not a sufficient requirement for assuring reliable predictions for a future world.

A number of studies has been carried out on multi-model skills. Solomon *et al.* (2007) and van der Linden and Mitchell (2009) provided general overviews of multi-model ensemble research. Santer *et al.* (2009) presented a multi-model detection and attribution study based on 22 different climate models, from which they identified simulated fingerprint patterns of anthropogenically caused changes in water vapour. They examined whether results were sensitive to model quality, and found the anthropogenic water vapour fingerprint to be insensitive to model uncertainties.

Collins (2007) explored a number of new approaches for dealing with a number of sources of uncertainty that arise in the prediction process. He discussed problems concerning the use of model ensembles in constructing probability density functions (PDFs). Frame *et al.* (2007) too discussed the use of ensemble-based 'probabilistic' climate forecasts, and pointed out a number of challenges related to experimental design and interpretation of forecast results. Knutti (2010) observed that the problem of weighting and evaluating models proves to be non-trivial, and suggested to downweight or eliminate a few very poor models that are clearly unable to mimic important processes, or are even physically implausible.

Murphy *et al.* (2007) presented a strategy for probabilistic predictions of future climate, based on a set of ensemble simulations of equilibrium and time-dependent changes. They designed a model ensemble by perturbing poorly constrained parameters controlling key physical and biogeochemical processes in a global climate model to allow quantification of the effects of uncertainties and internal climate variability on feedbacks. They also presented an ensemble of regional climate simulations at 25 km resolution over Europe, and proposed using a Bayesian statistical framework for generating probabilities constrained by observational metrics.

Maraun *et al.* (2010) reviewed validation methods and presented techniques for assessing downscaling skill. They paid special attention to the end-user needs in terms of skillful predictions, and discussed the models' ability to reproduce intensity-duration relations, spatial coherence, event size, statistical distribution of daily precipitation, and seasonal dependence of the downscaling skill. However, they did not mention the minimum skillful scales of the models, and they didn't really go into details about evaluating multi-model ensembles. Most studies on downscaling also involve selected limited regions or are restricted to a specific country, whereas a global representation is required to provide a representative picture of the global climate models (GCMs) ability to provide information useful for climate change adaptation, as regions may be susceptible to the GCMs' regional biases.

The criticism brought forward by Oreskes *et al.* (2010), Pielke Sr. and Wilby (2012), and Palmer (2011) is to some extent also related to the concept of *minimum skillful scale* which has been acknowledged in a number of studies (Benestad *et al.*, 2008; Huth and Kyselý, 2000; Zorita and von Storch, 1997; Grotch and MacCracken, 1991). Few of the evaluations of multi-model ensembles have addresed the question of minimum skillful scale, however. Grotch and MacCracken (1991) concluded that the quality of the model simulations of present climate sets one limitation in terms of projecting a future climate change. Based on Grotch and MacCracken (1991), Zorita and von Storch (1997) argued that the minimum scale and minimum skillful scales are not the same. The former should be defined as the distance between two neighbouring grid points, whereas the skillful scale is larger than n grid-point distances. Zorita and von Storch (1997) suggested that $n \geq 8$, which for a GCM with $2° \times 2°$ is about 16° (which corresponds to approximately 1780km at the equator, 1100km at 50°N/S, and 600km at 70°N/S). The presence of a skillful scale has also been acknowledged by Huth and Kyselý (2000), who also referred to the work by Grotch and MacCracken (1991). It is important to note that the minimum skillful scale may depend on the type of climate variable and time scale. It is also plausible that the concept of minimum skillful scales apply to the temporal dimension due to approximations made in solving complex equations, round-off errors, imperfect numerical algorithms, or the discrete representation of smooth functions.

Here the evaluation of climate models involves both global scales as well as their ability to describe regional and local conditions, down to the minimum spatial skillful scales. The paper is divided into two parts where the first involves an analysis of minimum skillful scales for a multi-model ensemble whereas the latter presents an evaluation of downscaled results on the scale useful for impact researchers and climate change adaptation. However, the data and methods are presented before discussing the results, and a discussion and the conclusion are provided after the results.

Data & Method

Here the evaluation of minimum skillful scale was carried out for the surface air temperature (2m), and involved 49 different simulations from the CMIP3 (Meehl *et al.*, 2007) data set (Table 1). The CMIP3 ensemble was chosen here rather than the more recent CMIP5 archive for several reasons: (a) they correspond to the downscaled results presented in Benestad (2011) that are evaluated in the second part of this paper; (b) it is useful to evaluate the CMIP3 in order to provide a reference against which the CMIP5 results subsequently can be evalued; (c) and the new CMIP5 archive is more cumbersome and impractical to get at whereas the CMIP3 data were conveniently at hand (limited resources slows down progress).

The NCEP/NCAR reanalysis (Kalnay *et al.*, 1996) was used for evaluating the global and regional mean trends of the CMIP3 ensemble as it was up-to-date and came with complete spatial coverage. Missing data gaps make the analysis complicated, since the same time-varying missing-data mask should be applied to all models. Care should be exercised when using reanalysis for trend analysis* (Hines *et al.*, 2000; Bengtsson *et al.*,

* http://climatedataguide.ucar.edu/

2004a), so two other gridded observations were included in the global mean comparison. Here the additional gridded observations were the GISTEMP (J.E. Hansen, 2009) and HadCRUT4 (Jones *et al.*, 2012; Morice *et al.*, 2012). The latter involved an ensemble of gridded results, and here the median was taken as best estimate and the 95% of the uncertainty estimates was taken from ensemble quantiles.

The trend evaluation for the downscaled results relied on the local temperature series taken from the Dutch Meteorological Institute's (KNMI) ClimateExplorer*, but regional mean trends were entirely based on the NCEP/NCAR reanalysis. In some regions, the gridded observations too will suffer from sparse network of data, and the 2-metre air temperature in the reanalyses is a variable that is constrained by observations through assimilations. Hence, both the reanalysis and the temperatures are less certain in areas with a low network density of thermometers.

Simulations from the 20th century were combined with corresponding run for the 21st, assuming the SRES A1b emission scenario. Annual mean values were extracted from the 1962–2011 interval based on the combined simulation results for the two centuries, and for the comparison with observations, anomalies were computed with respect to 1961–1991 baseline period. The total global mean temperature differed by 1.93 K between the different GCMs, suggesting an error range of 0.7 % in terms of the absolute values (the temperatures are calculated in degrees Kelvin). This error range is affected by e.g. the models' spatial resolution, the land-surface characteristic, parameterisation schemes, and cloudiness. The GCMs are usually "tuned" by varying some parameters used in a simplified description of unsolved processes (parameterisation schemes, e.g. describing clouds, etc) so that the model description gives a best match to the real world. Often, the criterion for tuning is the mean state and the annual cycle, and the tuning may

* http://climexp.knmi.nl/

have some effect on the climate sensitivity (Bender, 2008), however, the tuning does not involve fitting the transient integrations to the past temperature evolution.

Here 'trends' refer to the linear rate of change derived trough an ordinary least-squares regression using time as the only co-variate. Trends for the downscaled results were estimated for each of the four seasons (December–February; March–May; June–August; September–November). The length of the observed thermometer records varied from place to place, and the trend was therefore estimated for different time intervals for the different sites. However, in comparing the downscaled model results with observations, the trend was estimated over a corresponding interval for the downscaled results. Only station with more than 30 years of data were included in this evaluation exercise.

The simulated trends derived for the global and regional means, as well as for the downscaled results, were assumed to be approximately normally distributed with respect to ensemble member. Normal quantile-quantile plots were used to test this assumption for the global mean trends (Figure 1) and downscaled results (not shown). Figure 2 illustrates how best-fit normal distribution of the envelope of simulated global warming rates from CMIP3 (grey) compares with the corresponding trend estimated from the NCEP/NCAR reanalysis (black line). In this plot the text 'p= 28%' refers to the quantile (q_P) of the CMIP3 distribution that corresponds to the trend in the NCEP/NCAR reanalysis. Similarly, a quantile-quantile plot for the Svalbard temperatures (trends) indicated that the values of all the ensemble members were close to being normally distributed (not shown). Again, the observed trend was ranked against a best-fit Gaussian distributiuon to the downscaled CMIP ensemble. A high portion probability (rank) p around 50% suggests an under-confident ensemble whereas high frequencies of values near zero and 100% suggests over-confident ensembles (Weigel *et al.*, 2008).

Two types of evaluation were carried out to answer the questions: (a) if the GCMs reproduce the observed temperature trend statistics without emphasis on the precise location and (b) if the models were able to resolve the local temperature trend at a particular location.

The null-hypothesis of the first is that the CMIP3 results are statistically indistinguishable to the observations. In this case, skillful simulations are interpreted as the statistical distribution of the simulations being similar to that of the observations. This provides a test for whether the GCMs provide a reasonable description of the near surface trends on earth in general, but does not assess the question of whether the geographical distribution of trends is the same as in the observations. The test here is therefore to compare the range of values and to look for a universal distribution in the corresponding quantiles q_P.

The skill relevant for estimating the minimum skillful scale will be the extent to which the statistical characteristics of trend estimates projected for the past by the multi-model ensemble is distinguishable from corresponding observed trend. In other words, it is a question of whether the observed trend for a given location has the same characteristic as the corresponding ensemble predictions. Similar analysis was applied to the downscaled results in section 2. Hence, for the second question, one can apply standard evaluation techniques used in weather forecasting, such as the Receiver Operating Characteristic (ROC) and reliability plots (Wilks, 1995; Jolliffe and Stephenson, 2003).

The evaluation of global and regional mean trends only involved a few data points for area means, whereas robust estimation of ROC scores, reliability diagrams and BSS require larger data samples. However, these evaluation techniques were applied to the empirical-statistical downscaled (ESD) results with large number of individual sites ($N \sim O(1000)$). The ROC and reliability diagrams were derived using functions from the R-package `verification`*. The R-package was difficult to install because it dependended on other R-packages with deficiencies in the package structure. However, the specific R-functions needed for the ROC and reliability diagrams were extracted thanks to R's open-source philosophy, and were sucessfully complied on their own.

Simple schemes were used for evaluating the area mean trends. For the evaluation of the minimum skillful scales, the globe was divided into successively smaller regions by first dividing the world into 2 × 2 equal areas in terms of longitude × latitude sectors (2 × 2 gives 4 regions: 180°W–0°E/90°S–0°N, 180°W–0°E/0°N–90°N, 0°E–180°E/90°S–0°N, and 0°E–180°E/0°N–90°N). This process was iterated by subsequently dividing the globe into 4 × 4, 8 × 8, and 16 × 16 parts with the same longitude and latitude ranges (see sectors in Figure 3). Due to the earth's curvature, the regions near the poles represent smaller area than those with the same range of longitude in the tropics, however, this inhomogeneity is only present for the 8 × 8, and 16 × 16 regions.

The evaluation of area mean trends involved comparing the proportion of times the observations were outside a 90 percent interval (henceforth denoted ζ_{90}). About 10 percent of all the observed values is expected to fall outside this ζ_{90} range if the downscaled model results have the same statistical characteristics (location and variance) as the observations. The binomial distribution ($Pr(K = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$) provides a basis for judging whether the counts of trends outside the ζ_{90} range is anomalous, and to estimate the p-value. The binomial distribution was also used to evaluate the skill of the downscaled multi-model results in terms of describing extremely hot and cold seasons. About 5% of the downscaled results should on average fall below or above ζ_{90}.

The empirical-statistical downscaled results were the same as used in Benestad (2011), involving a common EOF frame-work based on large-scale 2-metre tempera-

[*] http://cran.r-project.org/web/packages/verification/index.html

ture. The calibration of the ESD models involved a stepwise multiple regression between common EOFs constructed from a combination of the ERA40 reanalysis (Simmons and Gibson, 2000) and individual GCM results. The data were de-trended and the climatology was subtracted prior to the calibration, and the common EOF frame-work ensured that the subsequent predictions used exactly the same spatial patterns as those used in the model calibration. Benestad (2011) evaluated the results by looking at R^2 from the regression between predictors and predictands in addition to a regression analysis between a set of best-fit coefficients of site-specific polynomial trend models to geographical parameters such as altitude. Another quality check was to compare the different trends for different calendar months, as it was expected that the trend would vary fairly smoothly with the seasons rather than abruptly and irregularly with spikes in some months. ESD results from GCMs with variance that differed by a half standard deviation from the observations were discarded from the downscaled ensemble. Using common EOFs in ESD involves some kind of quality control, but also ESD diagnostics such as inspection of spatial maps of regression weights and residuals give some indication about the quality of the results. Such diagnostics, however, only provide information about single GCMs. The objective of this study is to explore ways to evaluate the skill of the exercise of downscaling the entire multi-model ensemble.

Results

Skillful scale

The CMIP3 GCMs provide a reproduction of the global mean surface air temperature according to both the NCEP/NCAR reanalysis, GISTEMP, and HadCRUT4 (Figure 4). The model envelope matches well the observations as there was only two years (1967,

1996) when the annual mean was outside the 95.45% confidence interval about the ensemble mean (±2 standard deviations). According to a binomial distribution, the p-value associated with two such outliers is 60% and is well within the null-distribution (i.e. same distributions). The linear trend estimate for the global mean from the NCEP/NCAR reanalysis is furthermore well within the distribution of the CMIP3 ensemble (Figure 2), and the models therefore provide a skillful representation of the global warming.

The arguments presented by Oreskes *et al.* (2010) and Pielke Sr. and Wilby (2012), on the other hand, concern the question whether the GCMs are able to provide a skillful description of the temperature on smaller spatial scales. Figure 5 shows comparisons of the trend of the mean from a range of sub-domains, and there is a reasonable agreement between the models and the NCEP/NCAR reanalysis down to regions with approximately 20×10 grid-boxes, however the GCMs do not reproduce the strong warming associated with smaller spatial scales (panel a). A comparison between the area, independent of number of grid-boxes, suggests that model simulations of area mean temperature trends and observations are consistent on spatial scales of $10,000km^2$ and larger (panel b). The model results are inconsistent with the observations for spatial scales of $1,000km^2$ and smaller, however.

One interesting question is whether the minimum skillful scale depends on the number of grid-boxes or the area. A strong dependence on the number of grid-boxes may hint that the minimum skillful scale is related to the way the computer solves the equations through numerical schemes and discretisation. A strong area-dependence (both in terms of location and area size), on the other hand, may suggest that the computer solutions do not capture physical phenomena (e.g. the Arctic amplification). These aspects, however, may be entangled.

The various GCMs in the CMIP3 ensemble have different spatial resolution, implying that some of the regions in different models may cover similar areas but involve different number of grid-boxes. The results for the regional trends were displayed as coloured regions as in Figure 3, and it is evident that model-observations for the same area size could vary substantially from location to location. Some regions, notably over Europe, seemed to be more problematic more than others. One possibility may be that this exception is just a result of statistical fluctuations, although it is consistent with observations made by Oldenborgh *et al.* (2009): Europe warms faster than the models.

Skill of downscaled multi-model ensemble

Figure 6 presents a map of N sites where the 2-m temperature was downscaled. The colour shading indicates the length of the time series used for calibrating and evaluating the trends. Here the predicted trends were entirely based on the GCMs simulations of the 20th century, and didn't involve any information about trends from the ERA40 reanalysis, which was used only to train the model (on de-trended data).

Figure 7 shows the geographical distribution of the trend estimates, suggesting some degree of spatial coherence. The strongest warming is seen over Europe, and the picture is consistent with the observation with more pronounced warming over Europe than predicted by the GCMs (Oldenborgh *et al.*, 2009). Part of the evaluation of the downscaled results involved assessing whether model results and the observations belonged to the same statistical distribution. Figure 8 displays the distribution of trends for the four different seasons. It is clear from this comparison that the downscaled trends to some extend captures part of the observed range, but is not the same: the ESD results are biased and under-estimate the frequency of negative slopes.

To get more details about the discrepancies, the q_p corresponding to the observed trend was plotted in colour shadings at the various sites and for the four different seasons

(Figure 9). There was no clear geographical coherence in the ranking of the observed trend estimates among the corresponding ESD results, as there was for the observed warming trend (Figure 7). One exception may be over central Europe, however. A lack of spatial coherency may suggest that the model-observation differences is due to local data issues rather than the GCMs since the temperature from the GCMs is expected to be a smooth field in space. Histograms of the ranking statistics of the trend slope compared to the ESD results suggest that the ESD produces too conservative estimates (Figure 10): Often the observed trends slopes are steeper, be it in negative of positive sense. In other words, the downscaled CMIP3 ensemble is over-confident in terms of temperature trends. These results may partially be explained by the presence of pronounced secular internal chaotic variations, and an inspection of individual ESD, as presented in Benestad (2011), suggests that the CMIP3 ensemble do not always capture these internal secular variations very well (not shown).

The degree of under-confidence may potentially be explained in terms of the variane captured (R^2) by the regression analysis involved in the downscaling. Figure 11 shows a discernable relationship between the R^2-statistics and the ranking of trend according to the ESD estimates. This suggests that the mismatch is partly, but not wholly, due to reduced variance/poor fit between large and small scales.

It is also possible that short series are associated with greater scatter due to sampling fluctuations, in accordance with the expected greater trend estimates for shorter time series. Internal chaotic variations are expected to have increased influence of the trend estimates with shorter data records. A comparison between the length of the time series used for calibrating the ESD models and the trend estimates, reveals no systematic bias (Figure 12). The increased spread for shorter data records supports the notion that internal variations are present and do affect the estimation of the trend estimates by introducing extra noise and increasing the uncertainty. A comparison between Figures 6 and 8 suggests that the not all model-observation discrepancies are due to short series, as there are sites with divergent results even for the longest data records.

Both long and short series are associated with high and low R^2-statistics (upper panels in Figure 13). There might be a small hint of weaker regression results for the shorter series, however. The lower left panel in Figure 13 compares the quantile q_p of the ensemble distribution corresponding to observed values, by comparing the probability level p with the length of the record. The results suggests that the downscaled CMIP3 ensemble has a tendency to under-estimate the trends regardless of the length of the series.

Lower right panel in Figure 13 shows the proportion of the observed seasonal mean temperatures that falls outside the modelled ζ_{90}. For identically distributed data, one would expect to find approximately 10% of the data outside this 90% confidence range. Here, the results seem to cluster into a group with too few extreme values or too many. These results suggests that the downscaled results as a whole may not provide reliable information about the occurrence of heatwaves and cold seasons. For a small selection of sites, however, the downscaled results do produce results with similar statistical characteristics (Førland *et al.*, 2011), but we do not know if this is for the right reason. Discrepancies in trend characteristics in general may possibly be due to incorrect description of soil moisture, vegetation, boundary layers, blocking frequencies, or cloudiness.

The evaluation so far has indicated that the GCMs and ESD produce results with similar magnitudes as the observations, albeit with some biases in both the mean and the amplitudes. So far, the analysis has not addressed whether the ESD can provide a good description of how the results vary in space and with the geography. It is possible to use merhods from meteorological forecasts, such as ROC and reliability curves, to evaluate the spatial distribution of trends.

Figure 14 shows ROC plots for the four different seasons and for the predictions of local trends greater than 0.15°C/decade. Curves above the diagonal signify skill*, however, the curve for September–November follows the diagonal closely. These results suggests that the ESD results are associated with slight skill for most of the year, and the information from the analysis above indicates that the under-estimation of negative trends in the ESD results contribute to the low skill estimates.

The reliability diagram in Figure 15 indicates very low skill in predicting different geographical conditions in warming rates exceeding 0.15°C/decade. This evaluation suggests that the predictions have both minimal resolution and reliability, and they underestimate smaller probabilities while slightly overestimate the higher probabilities. The insert shows the relative frequency of predicted probabilities, and the peak at the low end may be attributed to sites with low R^2-scores. The downscaling of the CMIP3 models also tends to give a high proportion of cases with $dT/dt > 0.15$°C/decade, contrasting the the impression from Figure 8. However, the reliability diagrams assesses the predicted probabilities rather than the trend magnitude. The presence of internal variations in observations that are not equally present in GCMs leads to an over-estimation of the probabilities and still result in conservative estimates of the trend magnitudes.

Discussion

The results presented here may be interpreted as support for the scepticism against extracting local information from global climate models (Oreskes *et al.*, 2010; Pielke Sr. and Wilby, 2012; Palmer, 2011), however, the picture is more nuanced than presented in

[*] http://www.metoffice.gov.uk/research/areas/seasonal-to-decadal/gpc-outlooks/user-guide/interpret-roc

those papers for the predictive success (Hulme and Dessai, 2008). Although the present results only refer to one strategy for downscaling, the low skill in terms of predicting the geographical differences in warming rates reflect the GCMs' inability to describe regional climate just as well as shortcomings of the downscaling strategy. The analysis of the minimum skillful scale suggests that the former still is an important obstacle to getting reliable projections of local temperature. In itself, the evaluation of the GCM's ability to reproduce trends and their minimum skillful scales may also provide a basis that favours decision and learning success discussed by Hulme and Dessai (2008).

It is also important to keep in mind that the ESD results were statistically similar to the observed trend estimates in several respects, albeit with an under-estimation of negative trends. The statistical distributions of modelled and observed trends were distinct but the magnitudes were of similar order. The evaluation presented here revealed a fuzzy description of the geographical dependence of local temperatures, which suggests that the downscaled values may not capture the geographical distribution very well, under-estimate the internal variability and the most rapid warming rates, but nevertheless produce results that are approximately in the right ball park. It is interesting to note that a regression analysis against geographical indices[*] successfully identified high R^2 scores (>28% for Europe) for the ensemble mean polynomial trends (Benestad, 2011). While that analysis did not rely on observed trends (which may contain errors), they identified patterns consistent with real geographical features.

For many locations the trend is hard to estimate, as inter-annual variations give rise to a scatter in the trend-estimates. Furthermore, the trends in the past have been weak and much less pronounced than those expected for the future. Hence skillful reproduction of these past trends may require higher precision than is required for drawing useful

[*] altitude (z), distance to the coast (d), north–south distance, east–west distance, $\ln(d)$, $\sqrt{(}d)$ and $\sqrt{(}z)$

information from projected future trends of more substantial magnitudes. This is especially true in the presence of inter-annual and decadal variations. The results presented here suggest that CMIP3 under-estimate decadal variations. Furthermore, the internal variations are expected to be out-of-phase and influence the trend-estimates in different directions.

The ESD is designed to produce "semi-realistic" variance as it involves a regression against the observations (the variance predicted by regression models will always be the same or less than the original data), based on well-constained ERA40 temperature products. The geographic distribution of trend estimates are expected to be dependent on the year-to-year variance, unless the GCMs themselves predict regional trends that are incompatible with the observations. For ESD in isolation, any difference between the modelled variance can be due to model-observation differences, as the ESD used common EOFs as a basis (Benestad, 2001).

The analysis of the ESD results does not indicate region and continental-scale spatially coherent biases, but large variations within local sites as with the regional mean analysis. These large inter-site variations may mask GCM biases of larger scales, however. Important local factors such as landscape, land use, pollution, and urban heat island may not be sufficiently well captured by the GCMs, which could explain some of the discrepancies seen here. Furthermore, biases in the GCM results may not be time invariant, and discrepancies may also be connected with slight structural deficiencies in the models, for instance associated with the approximations, numerical or the parameterisation schemes. It is possible that all GCMs have common systematic errors.

It is also possible that errors in observations give the impression of lower skill than actually is the case. Here the temperature was taken from the KNMI ClimateExplorer, which may not incorporate the most recently quality controlled data. For some sites, the temperature record disagreed with reanalysis data (Benestad, 2011), and may look suspicious. For some African locations, the temperature exhibited conspicuously looking negative trends, but there was not sufficient information to rule out the question whether these were due to local environmental change or an introduction of irrigation. Furthermore, sparse distribution of measurements makes both homogeneity testing difficult and the reanalyses uncertain.

The ERA40 reanalysis used here to calibrate the ESD models may also have deficiencies in some regions (Benestad, 2011), possible due to shortcomings in the description of the hydrological cycle (Bengtsson *et al.*, 2004b). For the future, the analysis should be repeated with ESD models calibrated on newer reanalyses such as the ERA-Interim (Simmons *et al.*, 2007). Furthermore, a repeat on ESD based on the new CMIP5-ensemble will provide some information about added value associated with the next generation GCMs.

A central question is whether relevant processes and phenomena are well-captured by the GCMs, with a manifestation in realistic description of ENSO, NAO, storm track position and blocking frequencies. The CMIP3 models provide a crude description of such phenomena (Solomon *et al.*, 2007). Some models provide a better description of such phenomena than others, and in most cases, the phenomena reproduced in the models come with biases in geographical extent, time scales, or magnitude. For example, the results from some GCMs suggest too pronounced westerly air flow over Europe, with implications for both temperature and precipitation.

The results so far may suggest that there are indications of shortcomings, however, it is difficult to know how much of these can be attributed to model deficiencies and how much to imperfect observational records. In this respect, one should stress that observations are extremely important for the use of model results in real life applications, both in terms of model evaluations as well as providing additional and independent information about the real world. The scepticism towards using climate models for climate change adaptation in Oreskes *et al.* (2010); Pielke Sr. and Wilby (2012); Palmer (2011) is to a large extent neglecting the information from empirical data. For practical purposes, it is important to glean to historical trends and pose the question whether a shift in the trend is expected in terms of our physics-based knowledge.

The practical take-home message for end-users is that the downscaled temperatures tend to have realistic magnitudes for the trends albeit with a somewhat under-estimated range and missing accounts of cooling in some locations. It should be kept in mind that ESD may give a good description of the variance, even if the trends do not correspond to those observed in the past. The high precision needed to pin-point the exact trend estimate at particular location seems to be lacking in the empirical-statistical downscaled CMIP3 GCMs, possibly owing to the presence of secular internal variations, unaccounted for factors, shortcomings of the ESD, or possible inability to capture relevant phenomena. However, a lack of the high fidelity required to capture the weak trends of the past does not preclude the GCMs' ability to provide an approximate description of local temperature change in a world with a more pronounced global warming. The future development should nevertheless focus on the skill of the representation local phenomena in the GCMs, such as patterns in sea surface temperature, sea-ice, storm tracks, blocking frequencies, ENSO, the Monsoon, the MJO, and the mid-latitude westerlies. The misrepresentation of these may possibly be related to the models spatial resolution (Palmer, 2011).

Conclusion

Here it has been demonstrated that the CMIP3 GCMs are skillful at large spatial scales, and that there is a minimum skillful scale associated with the models' inability to represent smaller-scale phenomena. Nevertheless, by applying empirical-statistical downscaling to the multi-model GCM results for the 20th century simulations, one can obtain local warming rates for the past that on the whole are roughly of the same magnitude as observed. However, the downscaled results under-estimate the number of cases with negative temperature trends as well as the internal variability. The evaluation indicates that the downscaling is not able to provide reliable high-precision predictions of the geographical distribution of warming trends, but the usefulness of such scenarios will depend on the degree of precision required.

References

Bender, F.A-M. 2008. A note on the effect of GCM tuning on climate sensitivity. *Environmental Research Letters*, **3**(doi:10.1088/1748-9326/3/1/014001).

Benestad, R.E. 2001. A comparison between two empirical downscaling strategies. *Int. J. Climatology*, **21**, 1645–1668. DOI 10.1002/joc.703.

Benestad, R.E. 2011. A new global set of downscaled temperature scenarios. *Journal of Climate*, **24**(8), 2080–2098.

Benestad, R.E., Hanssen-Bauer, I., and Førland, E.J. 2007. An Evaluation of Statistical Models for Downscaling Precipitation and Their Ability to Capture Long-Term Trends. *International Journal of Climatology*, **27**(10.1002/joc.1421), 649–665.

Benestad, R.E., Chen, D., and Hanssen-Bauer, I. 2008. *Empirical-Statistical Downscaling*. Singapore: World Scientific Publishing.

Bengtsson, L., Hagemann, S., and Hodges, K.I. 2004a. Can climate trends be calculated from reanalysis data? *Journal of Geophysical Research*, **109**(doi:10.1029/2004JD004536).

Bengtsson, L., Hodges, K.I., and Hagemann, S. 2004b. Sensitivity of the ERA40 reanalysis to the observing system: determination of the global atmospheric circulation from reduced observations. *Tellus*, **56A**, 456–471.

Collins, M. 2007. Ensembles and probabilities: a new era in the prediction of climate change. *Philosophical Transactions of the Royal Society A*, **365**(doi:10.1098/rsta.2007.2068), 1957–1970.

Førland, E., Benestad, R.E., Hanssen-Bauer, I., Haugen, J.E., and Skaugen, T. Engen. 2011. Temperature and precipitation development at Svalbard 19002100. *Advances in Meteorology*, **2011**(Article ID 893790), 14.

Frame, D.J., Faull, N.E., Joshi, M.M., and Allen, M.R. 2007. Probabilistic climate forecasts and inductive problems. *Philosophical Transactions of the Royal Society A*, **365**(doi:10.1098/rsta.2007.2069), 1971–1992.

Grotch, S., and MacCracken, M. 1991. The use of general circulation models to predict regional climate change. *Journal of Climate*, **4**, 286–303.

Hines, K.M., Bromwich, D.H., and Marshall, G.J. 2000. Artificial Surface Pressure Trends in the NCEP-NCAR Reanalysis over the Southern Ocean and Antarctica. *Journal of Climate*, **13**, 3940–3952.

Hulme, M, and Dessai, S. 2008. Predicting, deciding, learning: can one evaluate the success of national climate scenarios? *Environmental Research Letters*, **3**(doi:10.1088/1748-9326/3/4/045013), 1–7.

Huth, R., and Kyselý, J. 2000. Constructing Site-Specific Climate Change Scenarios on a Monthly Scale. *Theor. Appl. Climatol.*, **66**, 13–27.

J.E. Hansen, NASA/GISS. 2009. *GISS Surface Temperature Analysis; Global Temperature Trends: 2008 Annual Summation*. http://data.giss.nasa.gov/gistemp/2008/.

Jolliffe, I.T., and Stephenson, D.B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester, England: Wiley. 254 pages.

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

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Wollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., and Joseph, D. 1996. The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**(March), 437–471.

Knutti, R. 2010. The end of model democracy? *Climatic Change*, **102**(doi: 10.1007/s10584-010-9800-2), 395404.

Maraun, D., Wetterhall, F., Chandler, R.E., Kendon, E.J., Widmann, M., Brienen, S., Rust, H.W., Sauter, T., Themeßl, M., Venema, V.K.C., Chun, K.P., Goodess, C.M., Jones, R.G., Onof, C., Vrac, M., and Thiele-Eich, I. 2010. Precipitation downscaling under climate change: Recent developements to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48**(2009RG000314).

Meehl, G.A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J.F.B., Stouffer, R.J., and Taylor, K.E. 2007. The WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. *Bull. Amer. Meteor. Soc.*, **88**, 1383–1394.

Morice, C.P., Kennedy, J.J., Rayner, N.A., and Jones, P.D. 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research*, **117**(doi:10.1029/2011JD017187).

Murphy, J.M., Booth, B.B.B., Collins, M., Harris, G.R., Sexton, D.M.H., and Webb, M.J. 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society A*, **365**(doi:10.1098/rsta.2007.2077), 1993–2028.

Oldenborgh, G.J. van, Drijfhout, S.S., van Ulden, A., Haarsma, R., Sterl, A., Severijns, C., Hazeleger, W., and Dijkstra, H. 2009. Western Europe is warming much faster than expected. *Climate of the Past*, **5**(1), 1–12.

Oreskes, N., Stainforth, D.A., and Smith, L.A. 2010. Adaptation to Global Warming: Do Climate Models Tell Us What We Need to Know? *Philosophy of Science*, **77**(0031-8248/2010/7705-0037).

Palmer, T. 2011. A CERN for climate change. *Physics World*, March, 14–15.

Pielke Sr., R.A., and Wilby, R.L. 2012. Regional climate downscaling - what's the point? *Eos*, **93**(5), 52–53. http://pielkeclimatesci.files.wordpress.com/2011/10/r-361.pdf.

Santer, B.D., Taylor, K.E., Gleckler, P.J., Bonfils, C., Barnett, T.P., Pierce, D.W., Wigley, T.M.L., Mears, C., Wentz, F.J., Brüggemann, W., Gillett, N.P., Klein, S.A., Solomon, S., Stott, P.A., and Wehner, M.F. 2009. Incorporating model quality information in climate change detection and attribution studies. *PNAS*, **106**(35), 1477814783.

Simmons, A, Uppala, S, Dee, D, and Kobayashi, S. 2007. *ERA-Interim: New ECMWF reanalysis products from 1989 onwards.* ECMWF Newsletter.

Simmons, A.J., and Gibson, J.K. 2000. *The ERA-40 Project Plan.* ERA-40 Project Report Series 1. ECMWF, www.ecmwf.int.

Solomon, S., Quin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignotr, M., and Miller, H.L. (eds). 2007. *Climate Change: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change.* United Kingdom and New York, NY, USA: Cambridge University Press.

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.* Met Office Hadley Centre, Exeter EX1 3PB, UK: European Comission.

Weigel, A. P., Liniger, M.A., and Appenzeller, C. 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Met. Society*, **134**(DOI: 10.1002/qj.210), 241260.

Wilks, D.S. 1995. *Statistical Methods in the Atmospheric Sciences.* Orlando, Florida, USA: Academic Press.

Zorita, E., and von Storch, H. 1997. *A survey of statistical downscaling results.* Tech. rept. 97/E/20. GKSS.

1 Figures and tables

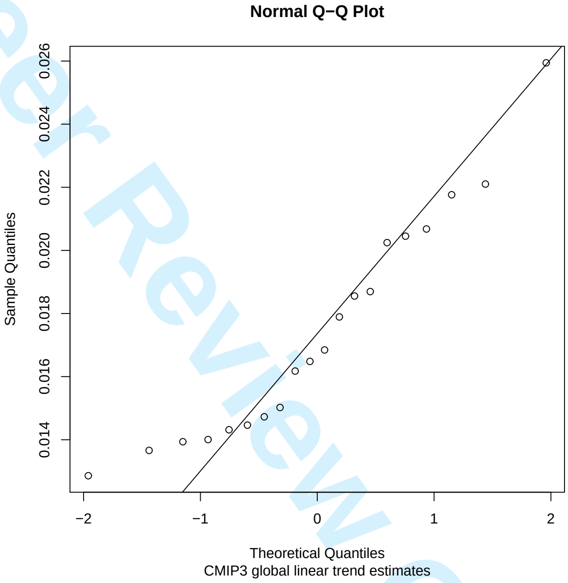

Figure 1. A normal quantile-quantile plot of the linear trend estimates (degrees/decade) for ensemble of CMIP3 simulations shown in Figure 4 suggests that the ensemble members are roughly normally distributed. The greatest deviation from the normal distribution is seen in the models producing the least warming.

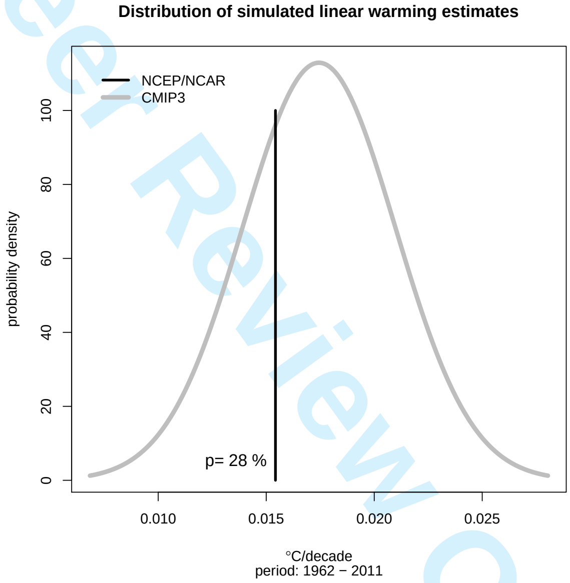

Figure 2. A comparison between the CMIP3 ensemble simulation and the NCEP/NCAR reanalysis estimate of the global mean linear warming trend. Each panel shows two sets of regions indicated by different hatching.

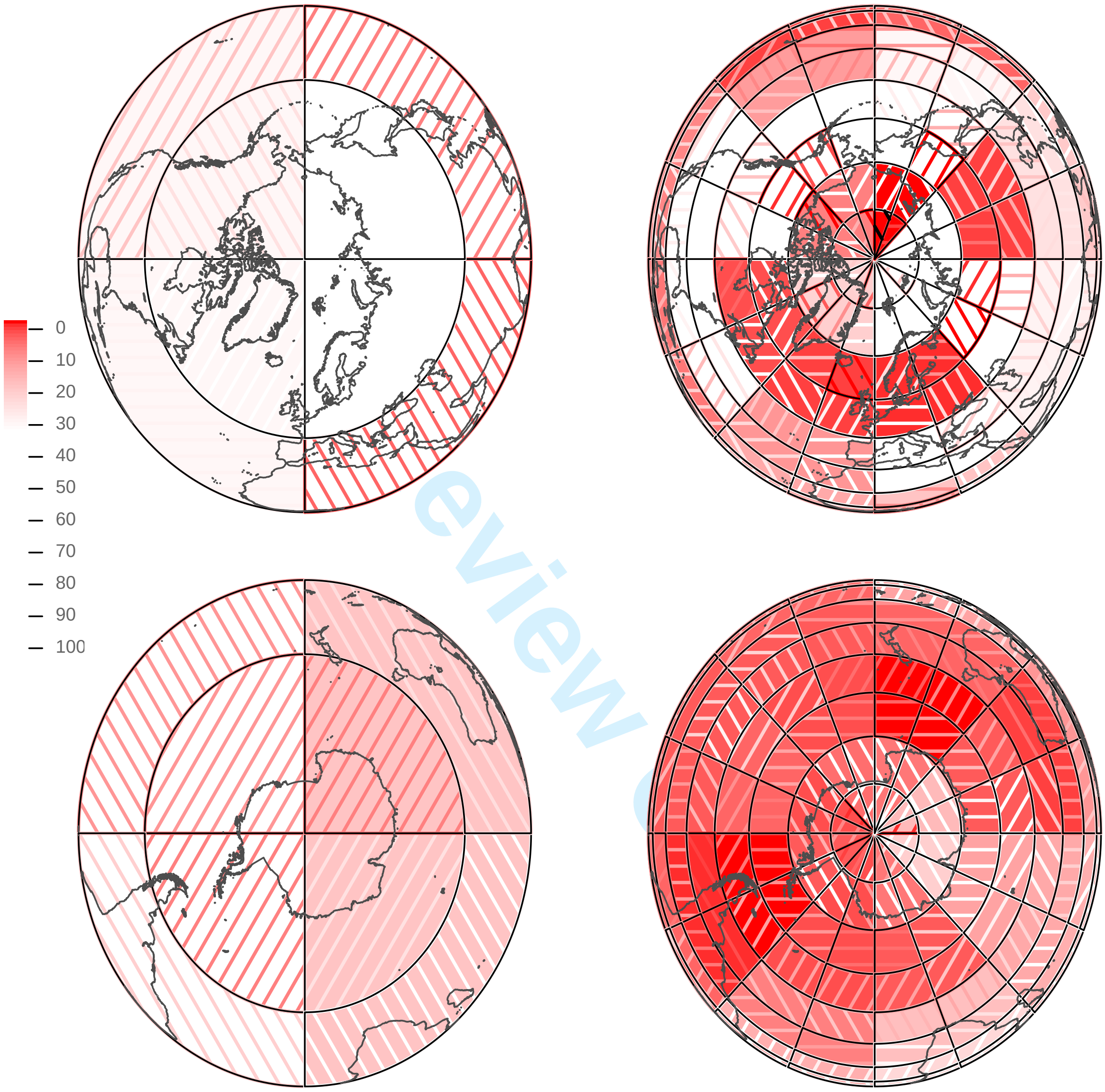

Figure 3. Map showing the regions and the quantile (shading) corresponding to the observed trend. The quantile was estimated assuming that the ensemble is normally distributed $\sim N(\overline{x}, \sigma)$.

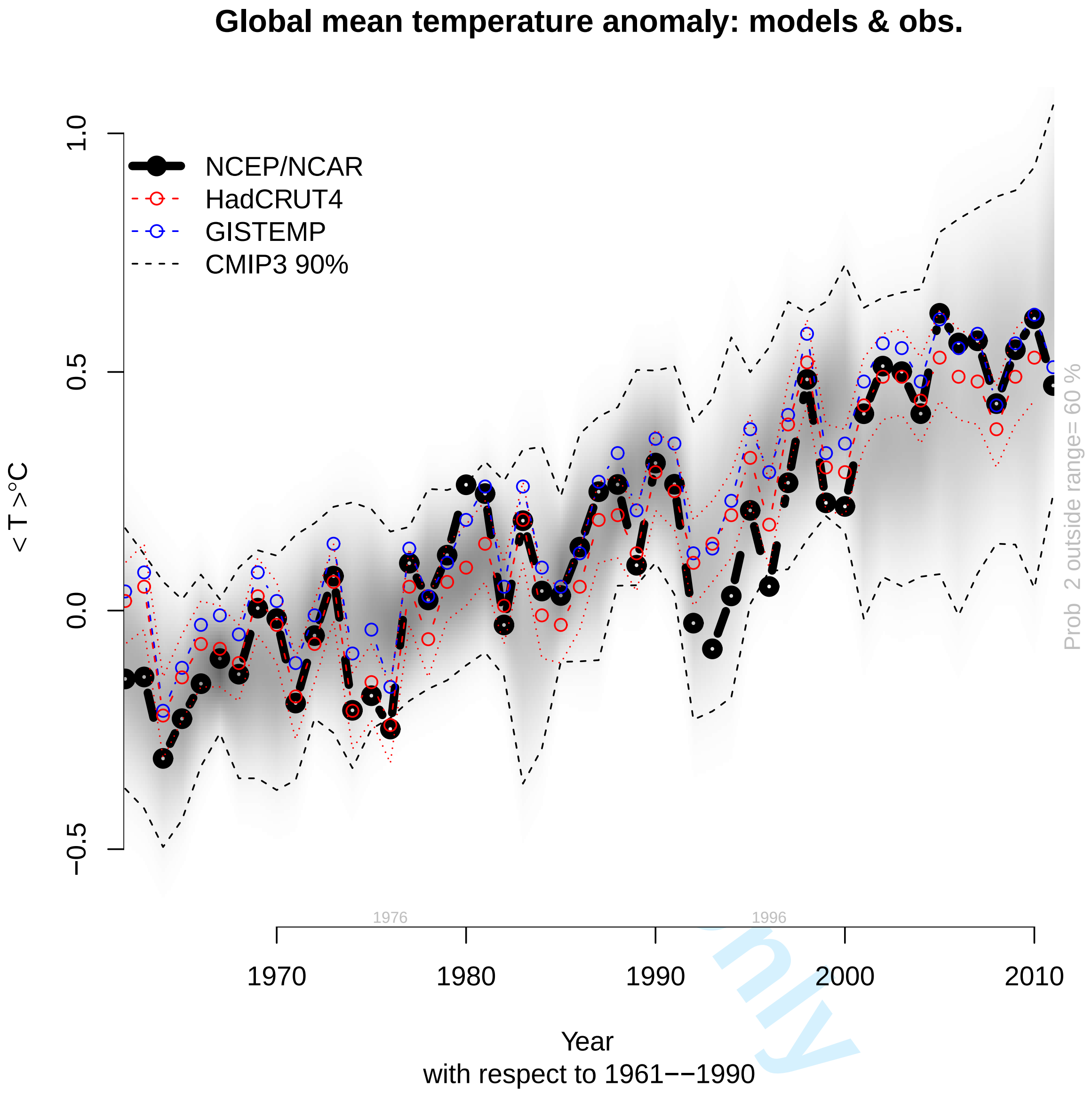

Figure 4. Comparison between the CMIP3 ensemble reproduction of the 1962–2011 annual mean temperature anomalies (wrt 1961-1990) and observed temperatures. The ensemble of simulations is shown as grey shading with darker shades for the central parts of the distributions. The thin dashed lines mark the ensemble mean ±2 standard deviations. The observations shown include the NCEP/NCAR reanalysis in addition to GISTEMP and HadCRUT4 (the dotted red lines indicate the 90% confidence interval for the HadCRUT4 analysis).

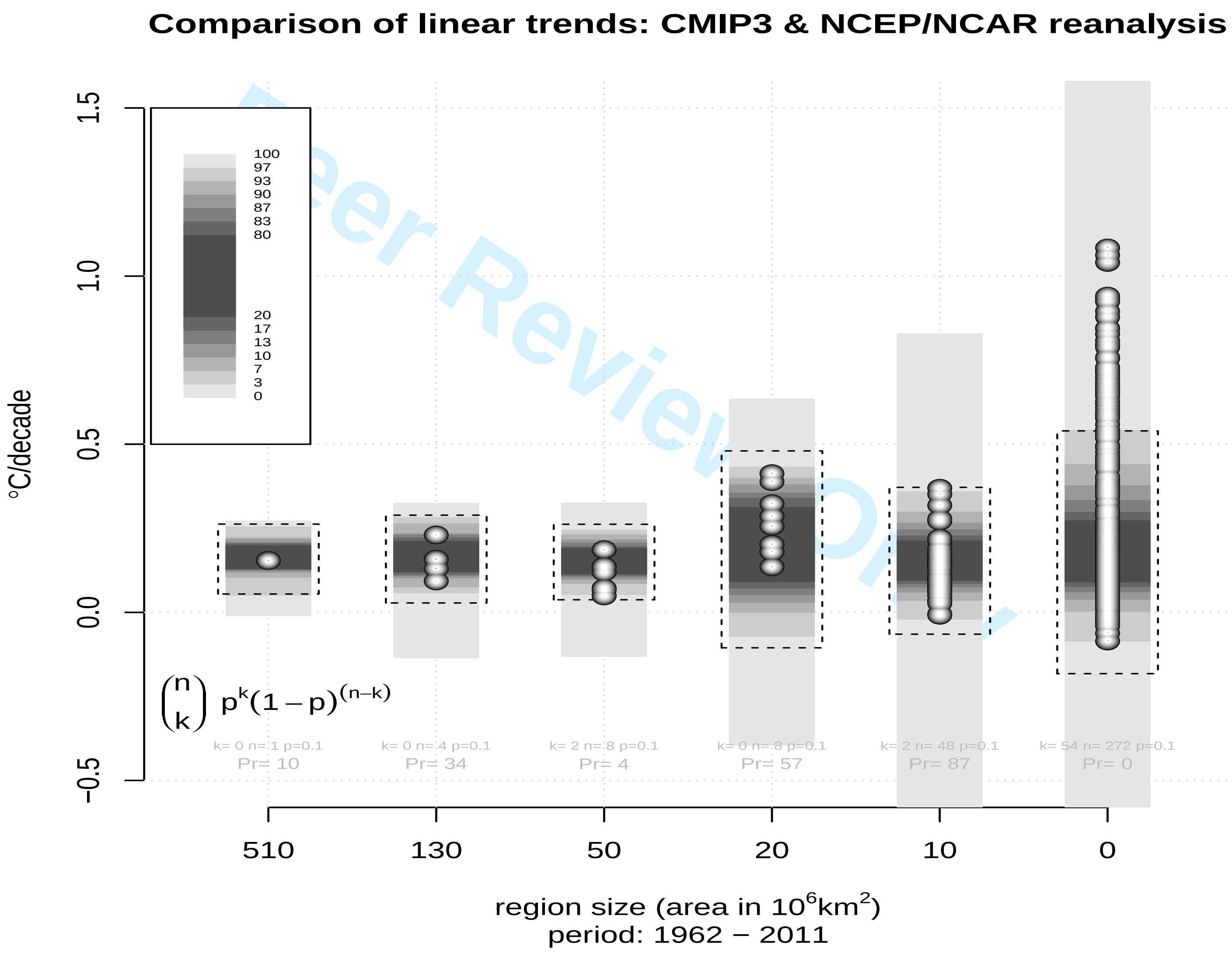

Figure 5. Comparison between regional trends simulated by the CMIP3 (shadings) and estimated by NCEP/NCAR reanalysis (symbols), sorted according to spatial scale. The models agree well with the reanalysis on the largest spatial scales, but the ensemble does not capture the stronger regional warming that has taken place on the smaller spatial scales smaller than $10^6 km^2$ despite a couple of outliers. The dashed boxes indicate the mean ±2 standard deviations.

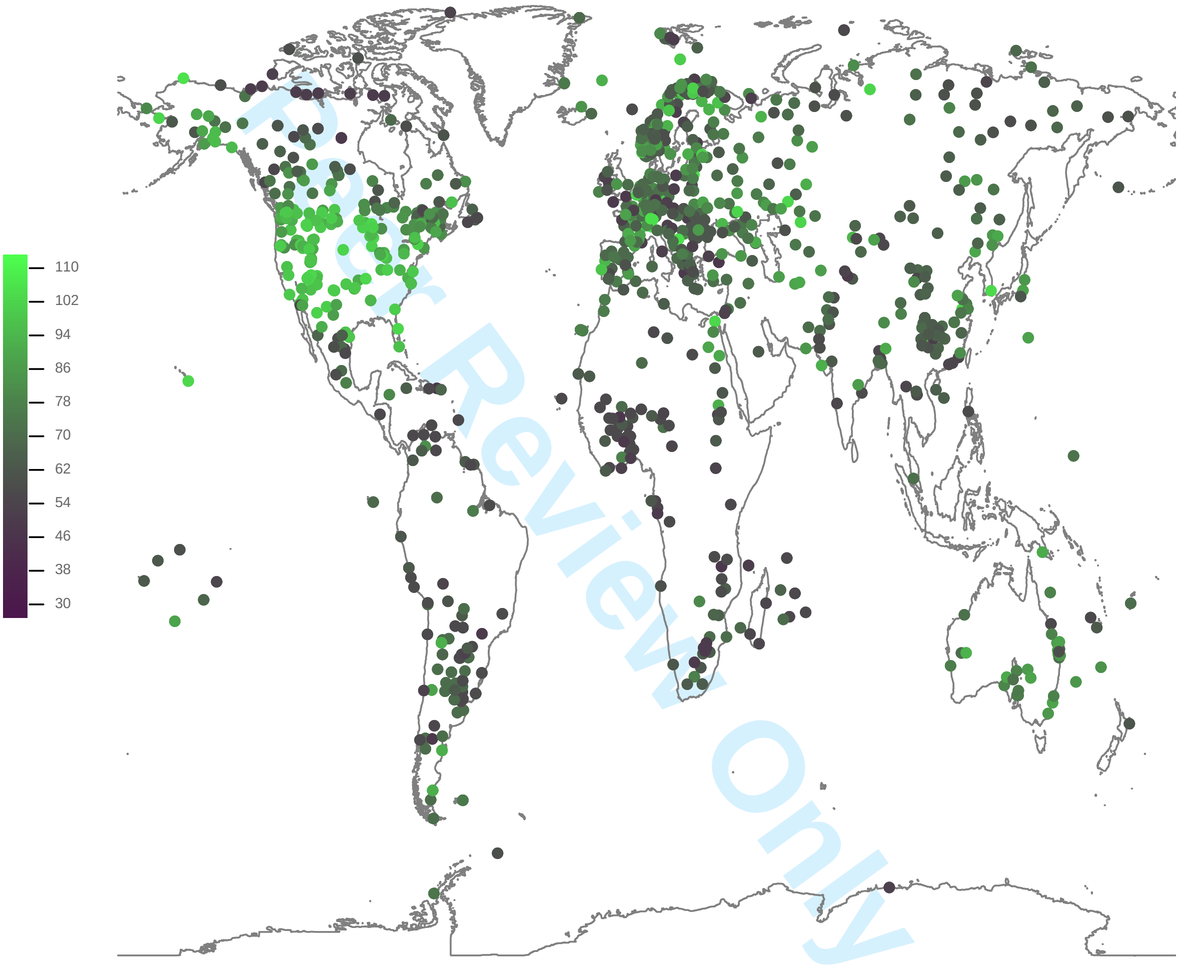

Figure 6. Overview of the station sites and the length of the data records (years). Light green suggest stations with more than 100 years of data, mostly found over the USA.

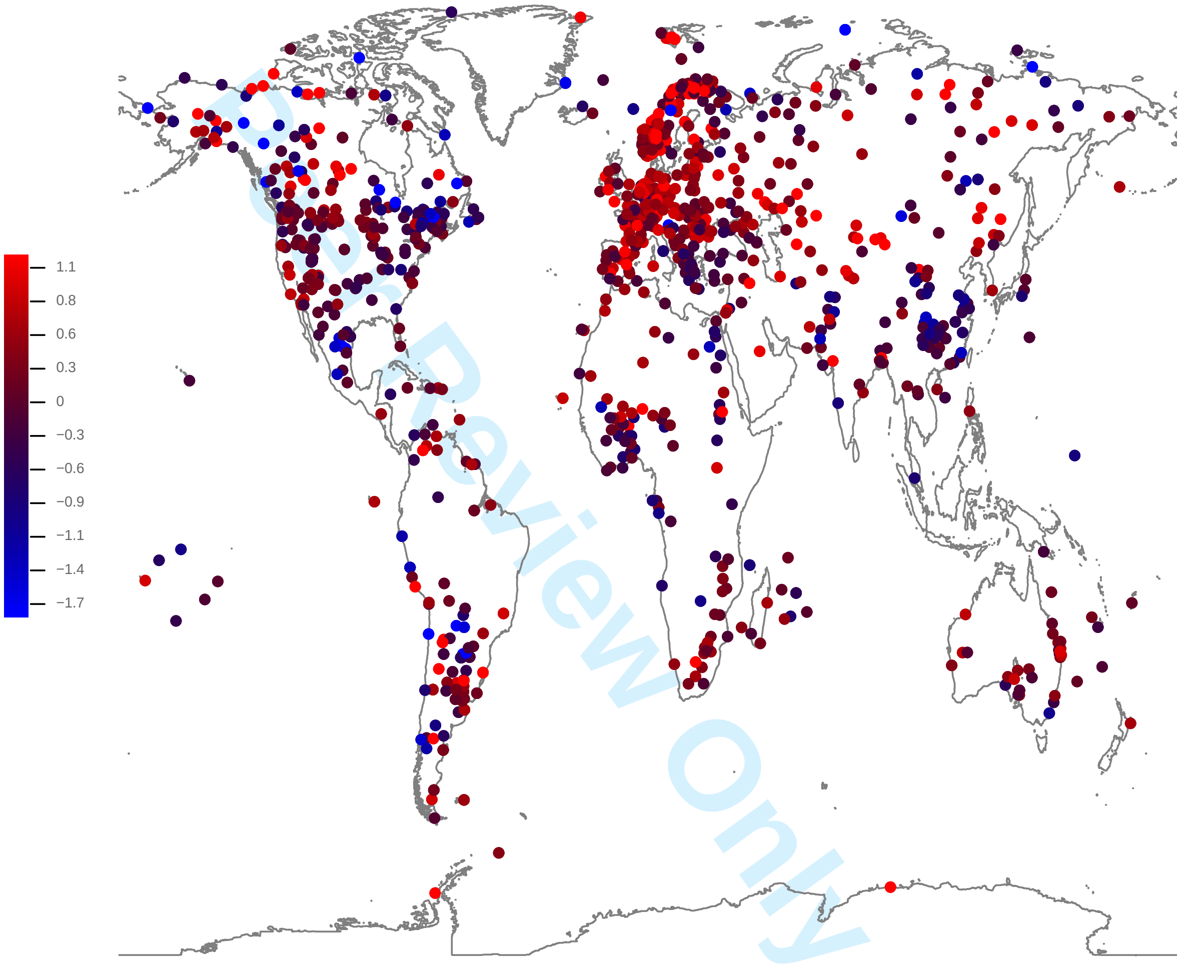

Figure 7. Overview of the station sites and the geographic distribution of trend estimates based on observed temperature.

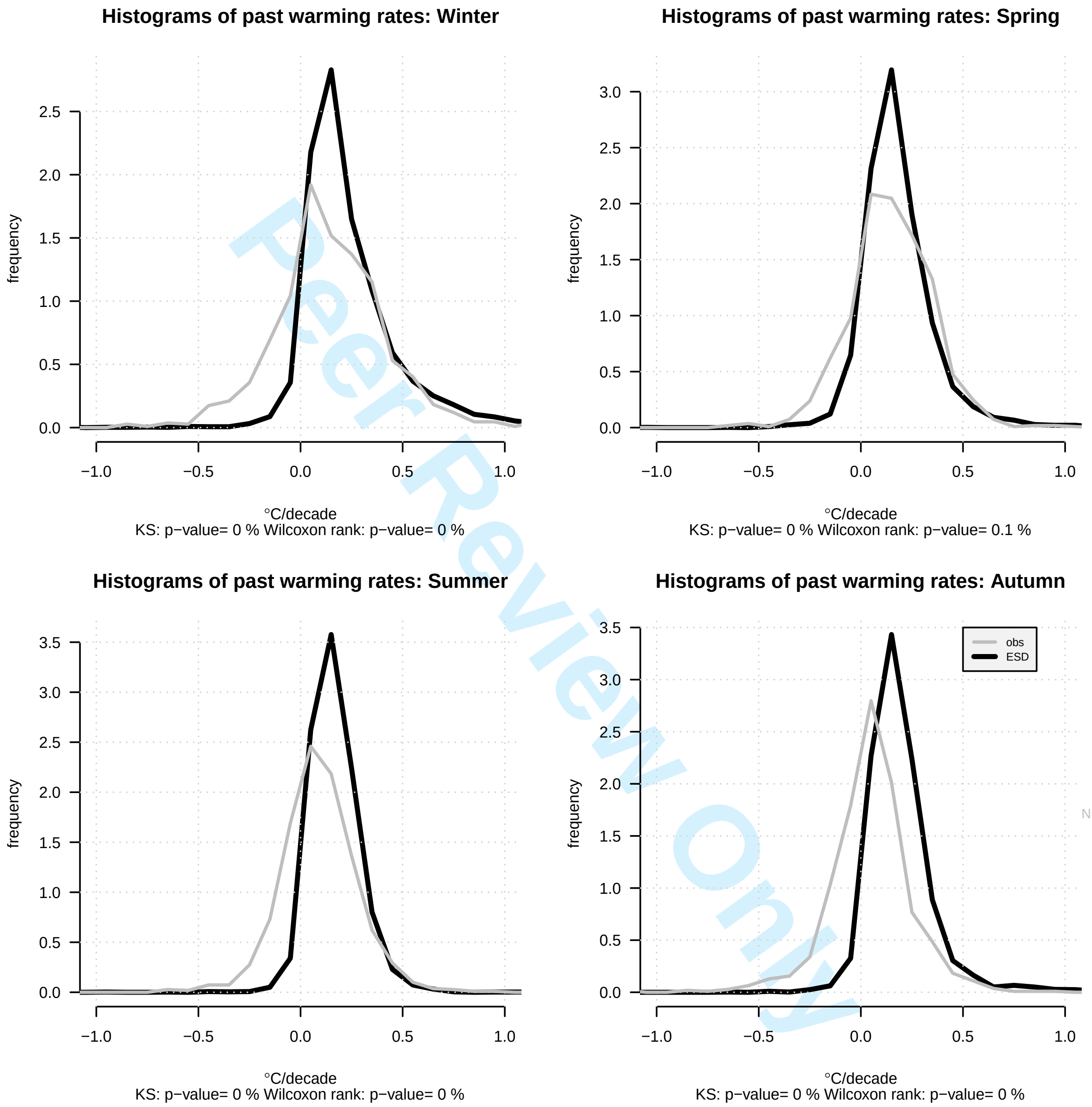

Figure 8. Histograms of observed (grey) and modelled (black) trend estimates for the four different seasons (different panels). In both cases, the trends indicate a positive bias (a global warming) compared to the observations, and the models tend to under-estimate the number of cases with negative trends (local cooling).

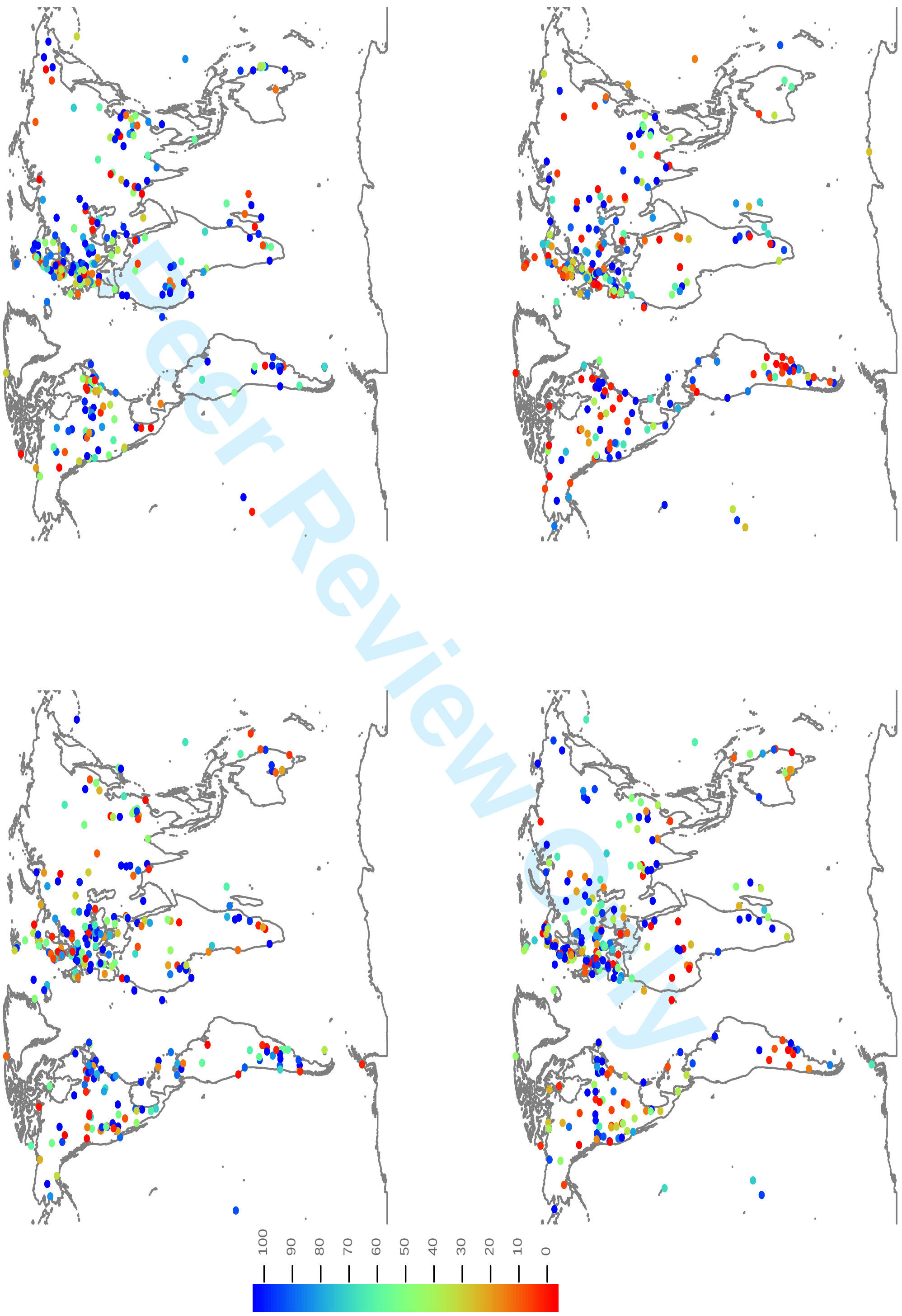

Figure 9. The geographical distribution of the quantile of the trend estimates simulated by the CMIP3 ensemble that corresponds to the magnitude of the observed trend. Red colour indicates an over-estimation of the historic trend whereas blue suggests an under-estimation.

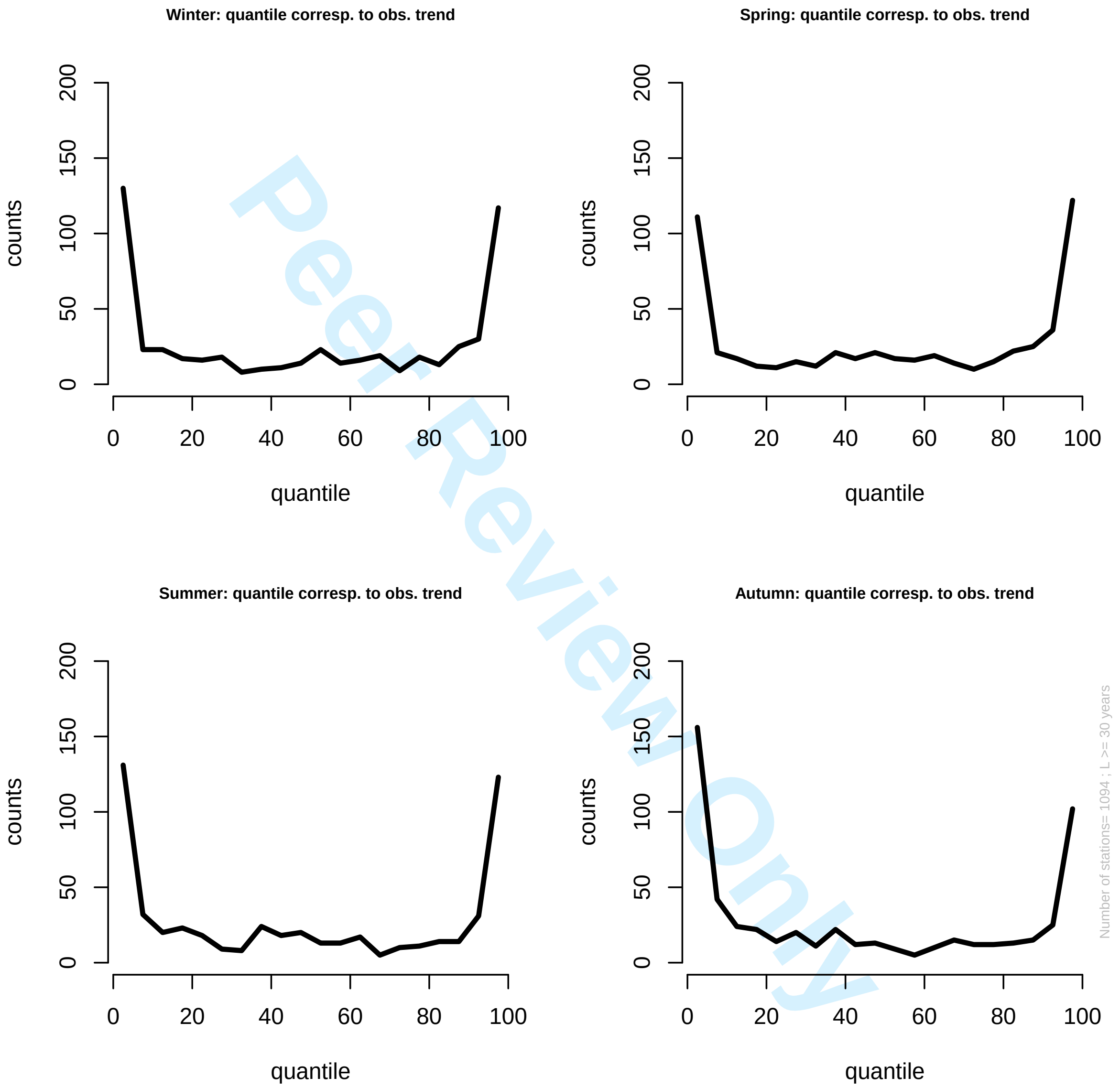

Figure 10. Histograms of the the quantile of the modelled trend estimates corresponding to the observed trend. The different panels show for different seasons, and all indicate that there is an over-representation of cases which are above or below the 90% C.I.

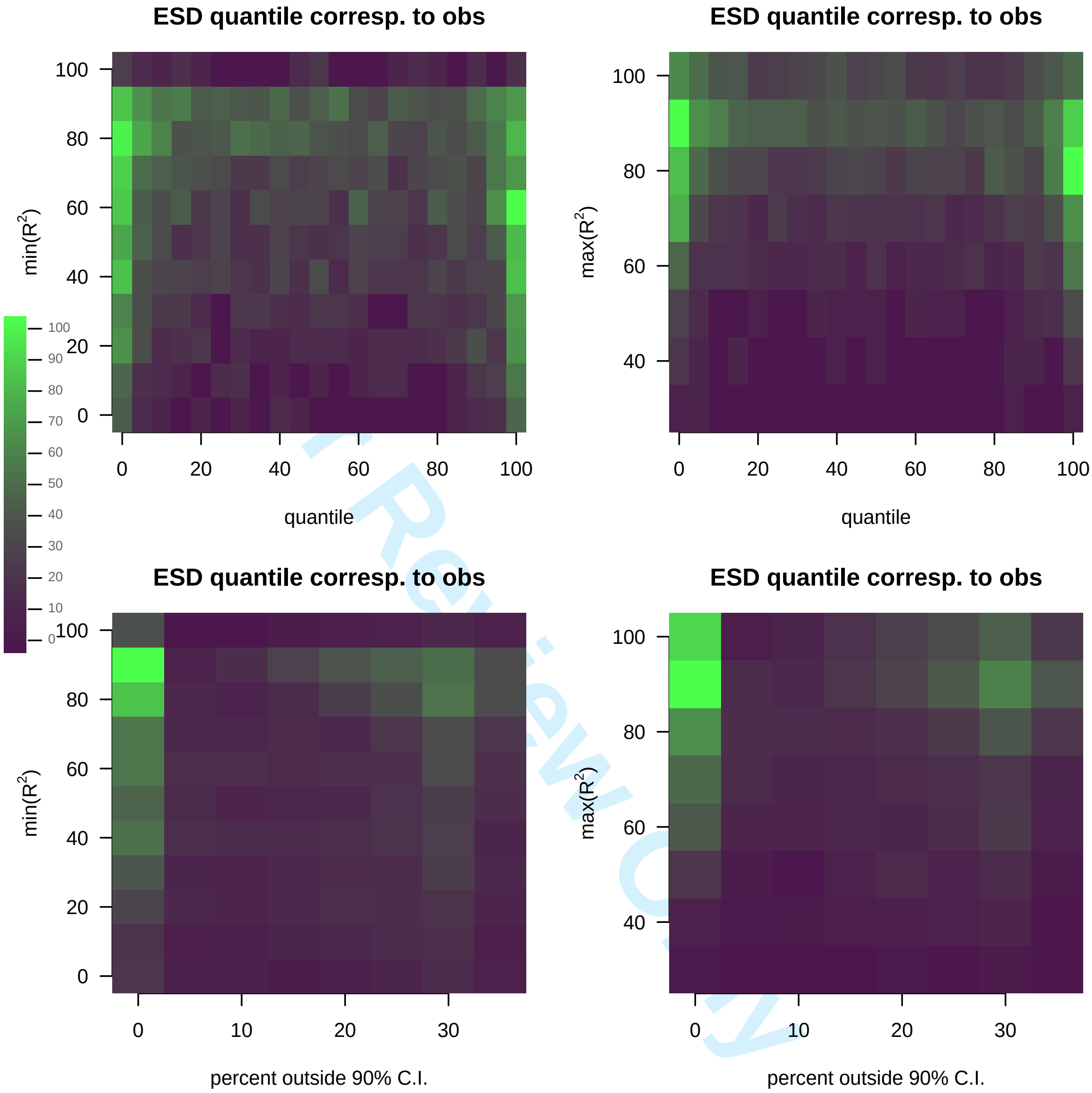

Figure 11. Correspondence between R^2 and the quantile of the modelled trend estimates corresponding to the observed trend. The colour shading indicate how the frequency of cases varies with minimum (left) or maximum (right) R^2 values and the quantile of the ensemble that corresponds to the observed trend (upper) and the proportion of the cases in which the observed trend is outside the 90% confidence interval simulated by the CMIP3 ensemble (lower).

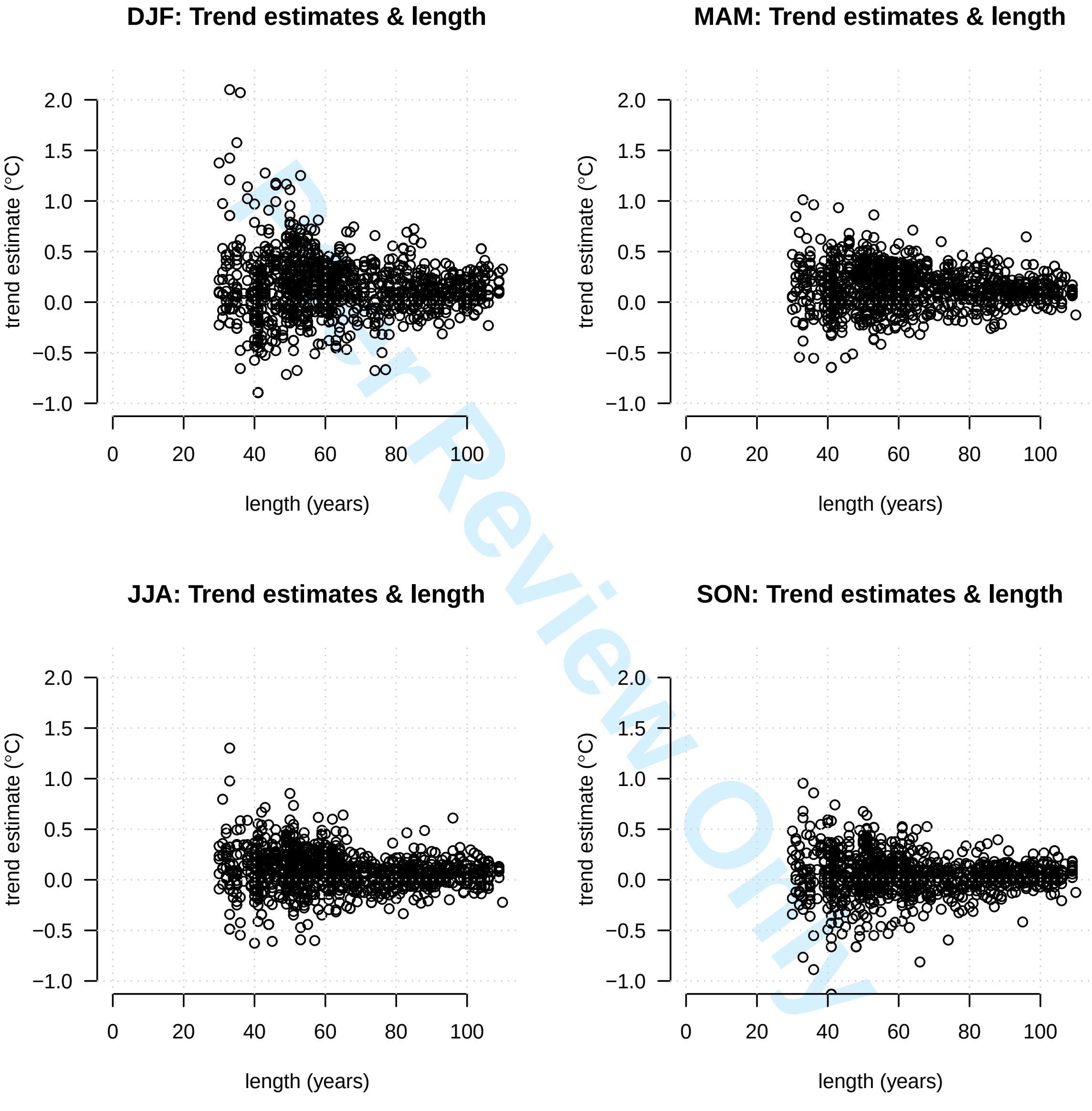

Figure 12. Correspondence between time series length and trend estimate indicates greater scatter for shorter series, but no systematic shift with respect to length. Furthermore, the high-estimates are all associated with short time series.

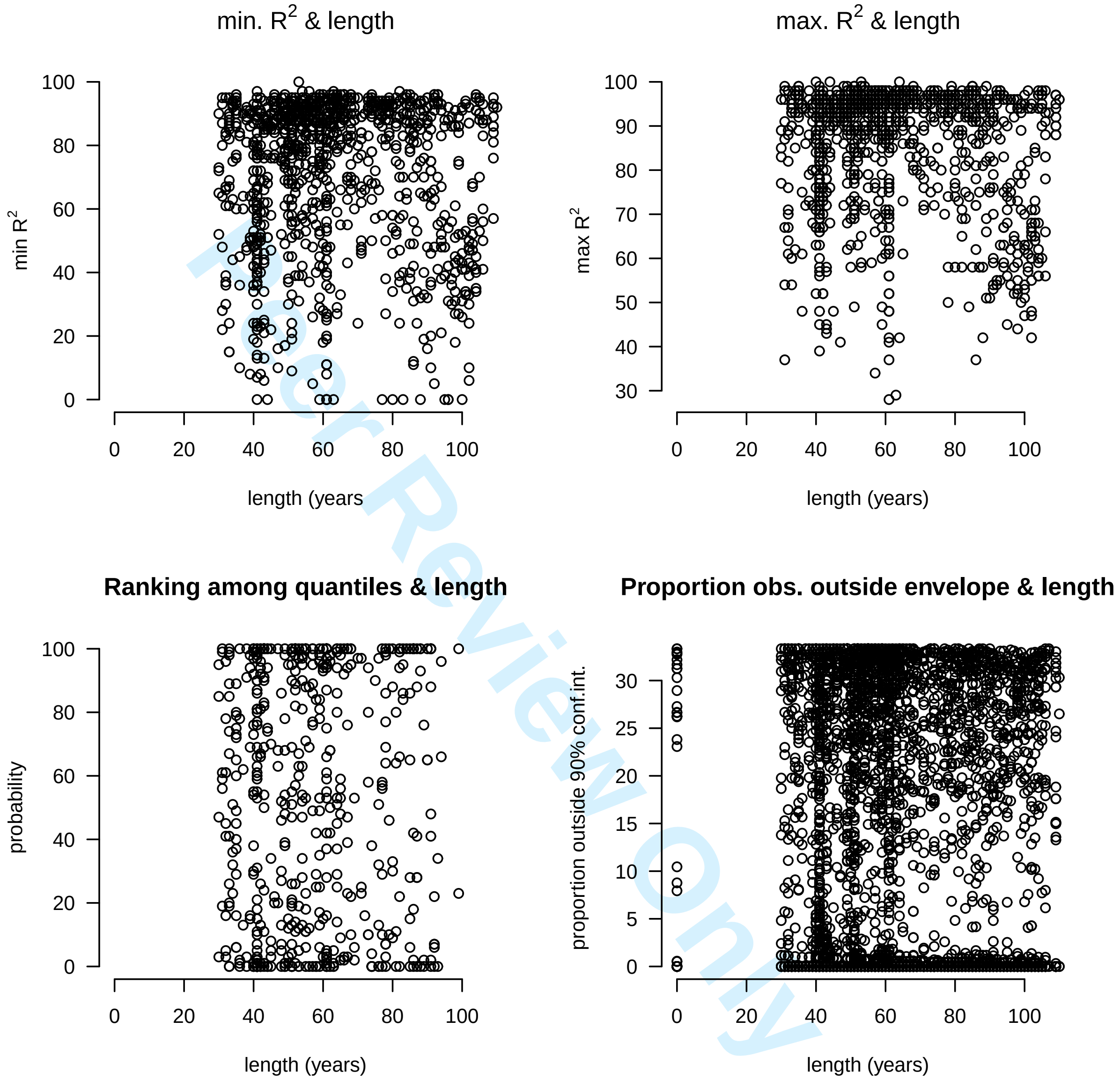

Figure 13. Correspondence between minimum (left) and maximum (right) R^2 and time series length (upper), and ranking of observed trends compared to the CMIP3 ensemble simulations (lower left). Lower right panel shows how the cases where observed trends are outside the 90% C.I. is distributed with respect the record length. Cut-off at 30 years, as stations with shorter records were not included here.

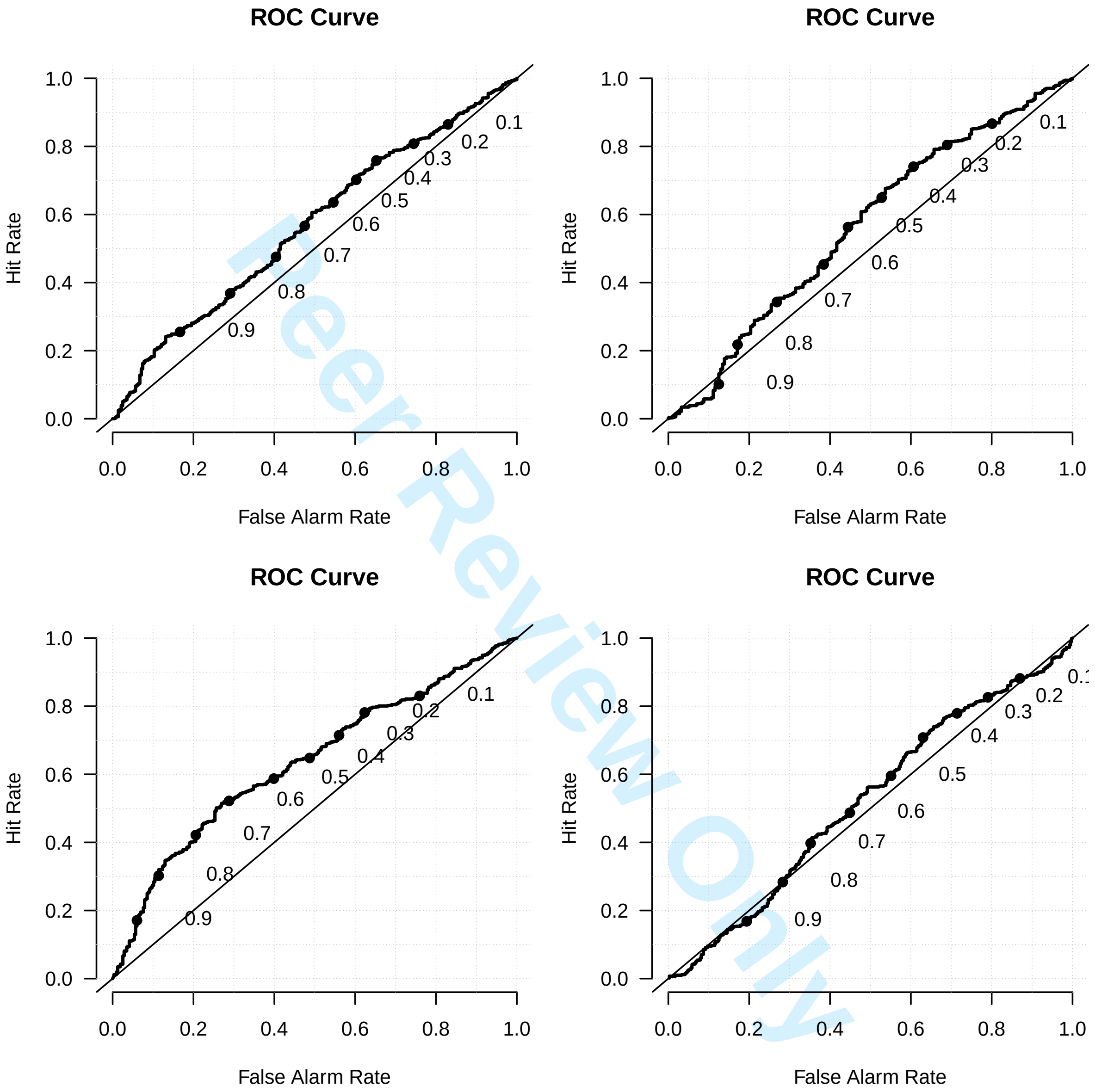

Figure 14. ROC plots for the ESD results over the 1963 locations and the four seasons are shown in the four different panels. The curves show comparisons between the hit rate and the false alarm rates for geographical distribution of past local warming exceeding a threshold value 0.15°C/decade. Best skill is seen for June–August, while the worst skills are seen during September–November.

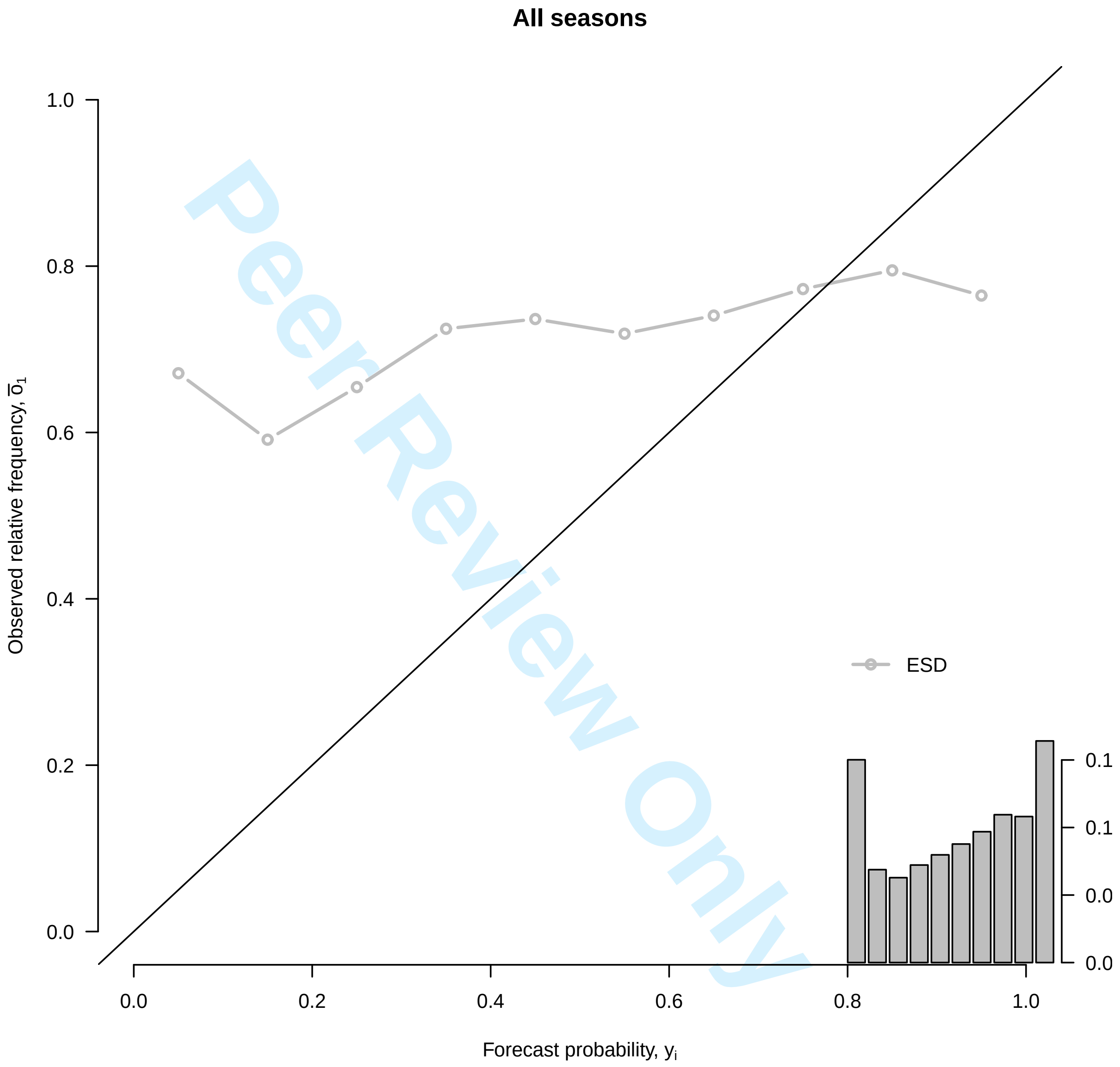

Figure 15. Reliability plot for all seasons and all locations showing a comparison between observed frequencies and corresponding forecasted probabilities of local warming trends exceeding a threshold of 0.15°C/decade. The insert shows the relative frequency of forecasted probabilities.

TABLE 1. List of the CMIP3 GCMs used in the analysis of skillful scales. The runs refer to both the '20c3m' simulations and the 'sresa1b', which were combined into on climate simulations for the time interval 1962–2011.

bccr_bcm2_0 - run1	cccma_cgcm3_1_t63 - run1	cccma_cgcm3_1 - run1
cnrm_cm3 - run1	csiro_mk3_0 - run1	gfdl_cm2_0 - run1
gfdl_cm2_1 - run1	giss_aom - run1	giss_aom - run2
giss_model_e_h - run1	giss_model_e_h - run2	giss_model_e_h - run3
giss_model_e_r - run1	giss_model_e_r - run2	giss_model_e_r - run3
giss_model_e_r - run4	giss_model_e_r - run5	iap_fgoals1_0_g - run1
iap_fgoals1_0_g - run2	iap_fgoals1_0_g - run3	ingv_echam4 - run1
inmcm3_0 - run1	ipsl_cm4 - run1	miroc3_2_hires - run1
miroc3_2_medres - run1	miroc3_2_medres - run2	miroc3_2_medres - run3
miub_echo_g - run1	miub_echo_g - run2	miub_echo_g - run3
mpi_echam5 - run1	mpi_echam5 - run2	mpi_echam5 - run3
mri_cgcm2_3_2a - run1	mri_cgcm2_3_2a - run2	mri_cgcm2_3_2a - run3
mri_cgcm2_3_2a - run4	mri_cgcm2_3_2a - run5	ncar_ccsm3_0 - run1
ncar_ccsm3_0 - run3	ncar_ccsm3_0 - run6	ncar_ccsm3_0 - run9
ncar_pcm1 - run2	ncar_pcm1 - run3	ukmo_hadcm3 - run1
ukmo_hadgem1 - run1		