JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 113, D21110, doi:10.1029/2008JD010100, 2008

Comparison of six methods for the interpolation of daily, European climate data

Nynke Hofstra
School of Geography, Oxford University Centre for the Environment, Oxford, UK

Malcolm Haylock
Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

Mark New
School of Geography, Oxford University Centre for the Environment, Oxford, UK

Phil Jones
Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

Christoph Frei
Federal Office of Meteorology and Climatology MeteoSwiss, Zurich, Switzerland

Abstract

[1] We compare versions of six interpolation methods for the interpolation of daily precipitation, mean, minimum and maximum temperature, and sea level pressure from station data over Europe from 1961 to 1990. The interpolation methods evaluated are global and local kriging, two versions of angular distance weighting, natural neighbor interpolation, regression, 2D and 3D thin plate splines, and conditional interpolation. We first evaluated, using station cross-validation and several skill scores, relative skill of each method at estimating point values, looking at spatial and temporal patterns and the frequency distribution of the variables. We then compared, for precipitation, gridded area averages from the candidate interpolation methods against existing high-resolution gridded data sets for the UK and the Alps, which are derived from a much denser network of stations. In both point and area-average cases, differences in skill between interpolation methods at any one point are smaller than the range in skill for a single method either across the domain, or in different seasons. The main control on spatial patterns of interpolation skill is density of the station network, with topographic complexity a compounding factor. The relative skill of different methods remains relatively constant through time, despite a varying station network. Skill in interpolating extreme events is lower than for average days, but relative skill of different methods remains the same. We select global kriging as the best performing method overall, for use in the development of a daily, high-resolution, long-term, European data set of climate variables as part of the EU funded ENSEMBLES project.

Received 10 March 2008; revised 25 June 2008; accepted 18 August 2008; published 11 November 2008.

Keywords: Interpolation, gridded daily climate data, Europe.

Index Terms:
0520 Computational Geophysics: Data analysis: algorithms and implementation; 3252 Mathematical Geophysics: Spatial analysis (0500); 3270 Mathematical Geophysics: Time series analysis (1872, 4277, 4475); 3309 Atmospheric Processes: Climatology (1616, 1620, 3305, 4215, 8408).

1. Introduction

[2] Gridded surface climate data are important for many applications, including climate change detection [e.g., *Barnett et al.*, 2005], the evaluation of climate models [e.g., *Caesar et al.*, 2006], the parameterization of stochastic weather generators [e.g., *Hutchinson*, 1995; *Price et al.*, 2000], and understanding how climate interacts with terrestrial, hydrological, and biogeochemical processes [e.g., *Goovaerts*, 2000; *Scholze et al.*, 2006; *Shen et al.*, 2001]. Current work in climate change modeling requires higher-resolution observational data for the evaluation of regional climate models. In addition to higher spatial resolution, daily temporal resolution is needed to evaluate the ability of models to simulate the variance and extremes in climate that are key for a range of climate impact assessments. While there are a wide range of regional and global data sets of monthly climate derived from either station and/or satellite observations, gridded daily data are either specific to a particular country [e.g., *Daly et al.*, 2002; *Hewitson and Crane*, 2005; *Perry and Hollis*, 2005], cover a short time period [e.g., *Piper and Stewart*, 1996; *Rubel et al.*, 2004], or have coarse spatial resolution [e.g., *Caesar et al.*, 2006]. Table 1
provides examples of available daily data sets; while not exhaustive, it is clear that there are no long-term, high-resolution, daily gridded data sets for Europe or other large regions.

[3] As part of the EU ENSEMBLES project a daily, high-resolution gridded climate data set for the European domain is required for evaluation of regional climate models developed within the project and for climate change impacts assessments. Three key sources for gridded data sets are meteorological station records, satellite observations and, for precipitation, estimates from weather radar. For surface air temperature and pressure, satellite observations are either not available (pressure) or have significant, spatiotemporal biases (temperature) that currently preclude their use in climatology and climate model evaluation. For precipitation, satellite and radar data have advantages (complete spatial coverage) and disadvantages (short temporal coverage and large biases) [e.g., *Gerstner and Heinemann*, 2008; *New et al.*, 2001; *Reynolds*, 1988]. Merged station-satellite data sets overcome some of these disadvantages, but remain limited in time duration, particularly at high spatial and temporal resolutions [*Huffman et al.*, 1995; *Kottek and Rubel*, 2007; *New et al.*, 2001]. A fourth source for gridded data sets are reanalysis data, such as ERA40 and NCEP/NCAR data sets. However, *Simmons et al.* [2004] show for temperature that these data are only comparable to observations from stations after 1979, and precipitation from reanalyses exhibit large errors and systematic biases.

[4] Given these issues with remotely sensed data, and the particular need within ENSEMBLES for gridded data extending back to the 1950s, the ENSEMBLES gridded data set has been based on interpolation of station data. As the gridded data are primarily aimed at regional climate model evaluation, the approach used aims to produce area average rather than point estimates. The methodology is described in detail by *Haylock et al.* [2008], but is a two stage process: station data are first interpolated as point estimates to a fine grid, after which the point estimates are averaged to obtain area averages for the 25 km and 50 km grids used by the regional modeling centers.

[5] The development of the gridded climate data set is dependent on both adequate underlying station observations and the use of an appropriate interpolation method for high-resolution gridded point estimates prior to creation of area-averages grid values. Members of the ENSEMBLES consortium have contributed an unprecedented number of daily station data for Europe as the basis for the resulting gridded data set [*Klok and Klein Tank*, 2008]. The objective of this paper is the comparison of several candidate interpolation methods for estimation of point data in order to identify the most appropriate method for use in the construction of the ENSEMBLES gridded data for precipitation, temperature and sea level pressure (SLP).

[6] We adopt two approaches for our comparison. First, we use station cross-validation [e.g., *New*, 1999; *Willmott and Matsuura*, 1995] over the European domain, where each station is excluded in turn and the station values are then estimated through interpolation from the surrounding stations. The measured and estimated values at the excluded station are then compared, enabling us to quantify the relative skill of different interpolation methods at estimating point values.

[7] Second, we wish to confirm whether the best performing method(s), identified through station cross-validation, are also the best when grid-box area averages are calculated from the point estimates. In the absence of true areal averages, we compare our gridded area-average estimates to area averages calculated from existing high-resolution gridded data for the UK and the Alps. These regional and national gridded data are based on an order of magnitude more stations than are available to the ENSEMBLES project. Our assumption is that these grids are a fair approximation to the true area averages, and those candidate interpolation methods that produce area averages from our sparser station networks that are closest to those calculated from the higher-quality grids will likewise produce better estimates of the true area average. Alternative approaches to evaluating area-average estimates include spatially sub-sampling from continuous fields such as weather radar or high-resolution regional climate model output, then interpolating from the sub-sample to estimate the original field. As the primary purpose of the ENSEMBLES data set is climate model evaluation, we wish to avoid comparing the interpolation methods on climate model fields. Radar-based precipitation estimates were considered, but we were unable to obtain these within the time-constraints of the research project.

[8] The paper begins with a review of previous work in the interpolation of daily station data (section 2), with emphasis on work that has evaluated different methods, and justification for the methods we choose to evaluate. This is followed by a description of the methods and data used in the evaluation (section 3). We then present our results (section 4), looking at relative interpolation skill from several different perspectives, and summarize our findings in section 5.

2. Interpolation Methods

[9] Many different interpolation methods have been used for the gridding of climate station data [see *Tveito et al.*, 2006 for a recent review]. According to *Vicente-Serrano et al.* [2003], the best performing interpolation method "varies as a function of the area and the spatial scale desired for mapping". Also important are the temporal duration and the nature of the climate variable to be interpolated; temperature and sea level pressure are, for example, more or less continuous in both time and space, whereas precipitation fields are spatially discontinuous on shorter timescales and more continuous on longer timescales [*New et al.*, 2001]. Moreover, the importance of geographical factors such as elevation [*Price et al.*, 2000; *Willmott and Robeson*, 1995], aspect, distance to coast [*Daly et al.*, 2002], seasonality and/or synoptic state [*Hewitson and Crane*, 2005], station density [*Willmott et al.*, 1994], and representation of the station network [*Willmott et al.*, 1991] may influence the choice of interpolation method and the accuracy of results. Some interpolation methods have the capability to include co-predictors, which may produce superior interpolation results. Another possibility is to reduce the influence of factors known to be important, through the interpolation of anomalies or through optimal interpolation. In the former, the deviation from the mean is interpolated before adding the anomaly field back onto a long-term mean field which is based on a much richer network of station means, often interpolated using co-predictors [*New et al.*, 2001; *Widmann and Bretherton*, 2000; *Willmott and Robeson*, 1995]. Optimal interpolation uses the long-term mean field as a "first guess" onto which shorter duration (monthly or daily) station values are interpolated [*Chen et al.*, 2002].

[10] *Tveito et al.* [2006]
note the importance of testing different interpolation methods for specific purposes. While numerous comparisons of interpolation schemes have been undertaken, these have mostly been for long-term mean data, or monthly [e.g., *New et al.*, 2000; *Price et al.*, 2000] or annual [e.g., *Vicente-Serrano et al.*, 2003] fields. In contrast, there have been few evaluations of alternative methods for interpolation of daily station data. *Kurtzman and Kadmon* [1999]
compare splines, and inverse distance weighting (IDW) and regression models, finding that a regression model predicts mean daily temperature values in Israel more accurately than splines or IDW. *Shen et al.* [2001]
qualitatively compare several methods for the interpolation of daily station data, concluding that most interpolation methods do not retain the variability of the data and over-smooth the raw station data, and are thus best for interpolating mean climate data, which are themselves much smoother. The number of days with precipitation is also often not adequately represented. *Shen et al.* [2001] adopted the nearest

station assignment, a hybrid of Thiessen polygons and IDW, as their best method [*Shen et al.*, 2001]. *Jarvis and Stuart* [2001] conclude that there is no significant difference between partial thin plate splines (TPS), ordinary kriging and IDW for daily minimum and maximum temperature values in England and Wales. *Daly* [2006] qualitatively compares IDW, TPS and versions of local and regional regression for the interpolation of precipitation, noting advantages and disadvantages of each method, but do not identify a "best" method. Finally, *Stahl et al.* [2006] compare, among others, Gradient plus Inverse-Distance-Squared (GIDS, a method based on multiple linear regression), ordinary kriging and IDW, finding that ordinary kriging performs best, except at high elevations, where GIDS performs best if the station density is high.

[11] From the above it is clear that different interpolation methods can work better for different variables, station densities and climate regimes. We therefore chose six different interpolation methods for evaluation of skill in interpolating the climate variables of interest: precipitation, SLP, and mean, minimum and maximum temperature.

[12] Natural neighbor interpolation (NNI), originally developed by *Sibson* [1981], is a fast and simple baseline method that has been used for many years as a standard part of the library of graphics functions provided by the National Center for Atmospheric Research (NCAR). NNI takes the best of Thiessen polygons and triangulation and objectively chooses the number of neighbors from which to interpolate based on the geometry. The weights for each station are selected based on the proportional area rather than distance. NNI produces an interpolated surface that has a continuous slope at all points, except at the original input points. It is an exact interpolator in that it reproduces the observations at the station locations.

[13] Angular distance weighting (ADW) has been used quite widely for interpolation of monthly climate data [e.g., *New et al.*, 2000; *Shepard*, 1984] and for daily data and extreme indices [*Alexander et al.*, 2006; *Caesar et al.*, 2006; *Frei and Schär*, 1998]. We test two versions of the formulation of ADW of *New et al.* [2000]. The first selects stations contributing to a grid-point estimate using a constant search radius of 250 km for precipitation and 500 km for temperature and SLP, with the distance components of the weights decaying to zero at the search radius. For the second version, which has only been applied to precipitation, the search radius and weighting function are permitted to vary across the grid domain, as explained by N. Hofstra and M. New (Spatial variability in correlation decay distance and influence on angular-distance weighting interpolation of daily precipitation over Europe, submitted to *International Journal of Climatology*, 2008). In both versions, if there are less than 3 stations within the search radius, the value for the grid point is not calculated; if more than ten stations are potentially available, the ten with the highest angular-distance weights are used.

[14] Conditional interpolation (CI) has recently been developed by *Hewitson and Crane* [2005]. Self-organizing maps (SOMs) are used to define characteristic synoptic rainfall states in a region surrounding the target grid point [*Hewitson and Crane*, 2002]. The search radius used for the SOM and the interpolation is set to 2.5 degree after a short sensitivity study. The synoptic state determines the likelihood of a wet/dry day. Wet day amounts are then interpolated using a weighted average of surrounding stations, where the station weights vary as a function of angular distance and are "conditional" on synoptic state. As CI was developed specifically for gridding precipitation, we do not evaluate CI for temperature and SLP.

[15] Regression has also been used fairly widely for interpolation, and can have various forms [e.g., *Stahl et al.*, 2006]. Here we test multiple linear regression using latitude, longitude, elevation and distance to coast as predictors. The regression relationship is established separately for each target point using only the neighboring stations within a radius of 500 km. Target locations with fewer than four neighbors are set to missing. We use singular value decomposition to calculate the regression coefficients because of its greater numerical stability than Gaussian elimination or LU decomposition [*Press et al.*, 1986]. Residuals are not interpolated separately.

[16] Kriging is used extensively in the geosciences [*Journel and Huijbregts*, 1978; *Kolmogorov*, 1939; *Krige*, 1951, 1966; *Matheron*, 1963]. It is a stochastic method that, like regression, applies best linear unbiased estimation (BLUE) methodology: the "estimated" (interpolated) value is a linear combination of the predictors (nearby stations), such that the sum of the predictor weights is 1 (unbiased) and the mean squared error of the residuals from the interpolating surface is minimized (best estimate). The interpolated surface is therefore a local function of the neighboring data, but conditional on the data obeying a particular model of the spatial variability (the variogram). Variogram modeling is done by fitting each of three non-linear functions (Gaussian, Spherical and Exponential) to the experimental variogram using the method of *Marquardt* [1963] and selecting the one with the lowest Chi-squared statistic. We tested for anisotropy but found the use of an isotropic variogram more appropriate. We also tried reducing the skewness of the precipitation distribution using a cube-root transform, but this did not improve the skill of the interpolation in cross-validation. The importance of elevation was tested by implementing this as an external drift, which improved the skill in temperature interpolation but not precipitation or SLP.

[17] Kriging is not an exact interpolator in that it will produce an interpolating surface that does not honor the observations at the station locations. We use two versions of universal kriging which differ in the manner in which the variogram is modeled: Global kriging (GK) where a single variogram is used across the entire region, for all days and stations and local kriging (LK) where a different variogram is defined for each interpolation point. We use the same variogram for each day of the year and found that defining separate variograms for each month reduced skill due to less data being available for variogram modeling. Search radii for stations were set, after comparing cross-validation skill scores for varying search radii, at 500 km for SLP, mean and minimum temperature, 450 km for precipitation and 300 km for maximum temperature. These differences in search radii do not influence the interpolated values significantly. For precipitation, we add an additional step, indicator kriging [*Deutsch and Journel*, 1998], where the occurrence of rainfall is interpolated (as binary values of 0 for <0.5 mm and 1 for >0.5 mm). The resulting interpolated values fall between 0 and 1 and can be interpreted as the probability of observing a wet day. Comparing the cross-validation skill of several probability thresholds revealed a threshold of 0.4 to be optimal for assigning a wet day to an interpolated point. Wet day amounts were then interpolated by universal kriging [*Webster and Oliver*, 2001], as for temperature and pressure.

[18] Thin plate splines (TPS) share features that are similar to kriging and there have been several comparisons of the two methods [*Hutchinson*, 1993; *Hutchinson and Gessler*, 1994; *Laslett*, 1994]. Splines use a different covariance function and one that is rarely used in kriging [*Hutchinson and Gessler*, 1994], which is defined by minimizing the generalized cross-validation error; thus, the amount of data smoothing can easily be optimized and TPSs are appropriate for use across large heterogeneous areas [*Hutchinson*, 1995]. Although there have been some attempts to unify the two approaches, such as the method of Matern Splines [*Handcock et al.*, 1994], the two methods are usually treated as independent. We use the ANUSPLIN package of *Hutchinson* [1993], comparing both 2D (TPS2D) and 3D (TPS3D) models. In the 3D implementation, elevation is converted to km and latitude and longitude are in degrees.

3. Data and Methodology

3.1. Station Data

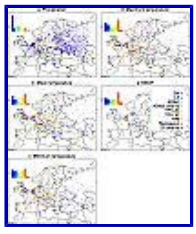

Figure 1. The interpolation method with the highest average ranking at each station, for each of the variables being assessed. Also shown is the frequency distribution of stations scoring best in each method.

[19] We use a preliminary version of a new pan-European data set of daily meteorological observations, collated by KNMI in collaboration with over 50 partners from European countries [*Klok and Klein Tank*, 2008]. We extracted records covering the period 1961–1990 from the larger data set (covering the period 1950–2006) for precipitation, SLP, and mean, maximum and minimum temperature. The station distribution (Figure 1) is particularly dense over the Netherlands, but is also good over Switzerland, Ukraine, Belarus, the Baltic States and Portugal. The distribution is less good over Poland, the Balkans, Scandinavia and Northern Africa (though many of these countries have contributed additional data for the final data set). The data have been quality controlled, so potentially erroneous outliers have been removed and potential inhomogeneities in station time series have been flagged [*Klok and Klein Tank*, 2008].

[20] For all variables, except for the CI method, data are first converted to anomalies relative to the long-term (1961–1990) monthly mean. Because of the nature of the CI method, data are not converted to anomalies. For the resulting comparisons, all interpolated anomalies are converted back to actual values by adding the long-term monthly values for the station time series. Rainfall anomalies are calculated by dividing the daily rainfall by the monthly total and therefore represent the proportion of the monthly total falling on that day. Anomalies have been used for all steps of the interpolation methods; for instance, GK and LK variogram modeling was done just using the anomalies.

3.2. Alternative Gridded Data Sets

[21] The UK precipitation data set is gridded at 5 × 5 km and based on a network of 4400 stations [*Perry and Hollis*, 2005], and is available for the entire evaluation period, 1961–1990. The station data are interpolated using multiple regression with geographic factors as the independent variables, followed by inverse distance weighting (IDW) of the residuals. The Alpine gridded analysis [*Frei and Schär*, 1998] covers a 700 × 1200 km^2
area including the Alpine mountain range and adjacent foreland areas. The analysis is based on more than 6500 stations from the high-resolution rain-gauge networks in the area. A modified version of the Shepard algorithm [an ADW technique, *Frei and Schär*, 1998; *Shepard*, 1984] was adopted for the interpolation of relative anomalies from the monthly climatological mean. The long-term climatology was derived with a local regression approach [PRISM, *Daly et al.*, 2002] specifically calibrated for the Alps [*Schwarb et al.*, 2001]. The procedures for the Alpine data set are exactly as in *Frei et al.* [2006], except for the grid resolution which is higher in this paper (0.25 × 0.167°, approx. 18 km). Our comparison is restricted to the period with maximum station density (1966–1990) with some additional constraints for Austria, where good quality station data are available for a shorter period (1971–1990). For our comparison, we wish to evaluate interpolation skill at 0.25° resolution, the finer of the two standard ENSEMBLES grid resolutions, so we regrid both of these higher-resolution data sets to 0.25°, through area averaging. The interpolated data have been developed by an interpolation to the locations of the grid points of the existing data sets using the two interpolation methods that performed best in the cross-validation. These data have then also been averaged over the 0.25° grid.

3.3. Skill Scores

[22] We use several skill scores to quantify both average skill and skill in interpolating extreme values. For precipitation, we also evaluate skill at capturing state (wet or dry). Different skill scores highlight different features of the results. For example the root mean-squared error emphasizes high errors, while the mean absolute error is less sensitive to extreme values [*Vicente-Serrano et al.*, 2003]. Therefore we have chosen to use five skill scores for the amount and two further scores for precipitation state. These are summarized below and their formulation is described in Table 2. Note that these skill scores are not error estimates for the final ENSEMBLES gridded data, which are a function of several factors, as discussed by *Haylock et al.* [2008]; here we are simply interested in relative skill at producing point estimates.

[23] 1. Compound relative error (CRE) is a measure of similarity between the interpolated and observed values. The correspondence is measured in terms of relative departures from the mean and

means and absolute variances of the two series. It is bounded below by 0 (best case) and unbounded above. A disadvantage of the CRE is that this skill score tends to favor interpolations that are too smooth. The method is sensitive to outliers [*Murphy and Epstein*, 1989; *Schmidli et al.*, 2001].

[24] 2. Mean absolute error (MAE) is a natural, unambiguous, measure of average error [*Willmott and Matsuura*, 2006]. It shows the errors in the same unit as the climate variable itself. MAE is bounded below by 0 (best case) and unbounded above.

[25] 3. Root mean-squared error (RMSE) is very commonly used as a measure of deviation from the observed value. Although is has been criticized as being ambiguous [*Willmott and Matsuura*, 2006] and its dependence on the squared error means that it is not resistant to outliers deviating from a Gaussian distribution. We include it because of its familiarity and, like CRE, its sensitivity to large outliers.

[26] 4. Linear error in probability space (LEPS) is a skill measure used extensively in seasonal forecasting [*Ward and Folland*, 1991]. It measures the error in probability space with reference to climatology. Like RMSE it indicates any deviation from the observed value. Unlike RMSE it is resistant to outliers as it assumes no particular statistical distribution. LEPS takes the values −1 (worst case) to 1 (best case). A value of 0 indicates the same skill as predicting the median for every case.

[27] 5. Correlation coefficient (R) also depends on squared deviations and so is similarly not a resistant measure. However, this statistic removes the effect of any bias in the interpolated data and highlights just problems with modeling the daily variability. Problems with correctly capturing the variance will not be highlighted as the measure normalizes the observed and modeled values by their standard deviations. The statistic is standardized and, therefore, can be compared across regions and months. However, because of its insensitivity to biases and errors in variance, the correlation coefficient should be considered as a measure of potential skill [*Murphy and Epstein*, 1989; *Wilks*, 2006].

[28] 6. Proportion correctly predicted (PC) measures the fraction of correct predictions. We use this measure to determine whether the state of precipitation (wet/dry) has been predicted correctly. While PC is a direct and intuitive measure of accuracy, it may not be the best measure when one of the events is less common than the other, as in precipitation wet/dry days [*Wilks*, 2006]. PC is bounded by 0 (worst case) and 1 (best case).

[29] 7. Critical success index (CSI) is a more appropriate method to determine accuracy when one of the correct events is less common than the other. It can be calculated by dividing the amount of correctly predicted least common event by itself plus the amounts of incorrectly predicted events (both most and least common). We use CSI to determine whether (1) the state of precipitation and (2) extreme events have been predicted correctly. For the latter we define a station value as extreme if it falls below the 5th percentile or above the 95th percentile; estimated values are defined in the same way. A correct prediction occurs when both the observed and estimated values satisfy these criteria, and we calculate a separate CSI for low and high extremes (CSI-low and CSI-high). For precipitation, we only calculate CSI-high. CSI attains a value of 1 for a perfect interpolation and a value of 0 for an unbiased random interpolation [e.g., *Wilks*, 2006].

[30] The skill scores are calculated for individual stations and as areal averages, for the full 30 year period, for different seasons (summer and winter).

4. Results

4.1. Average Cross-Validation Skill Over the Domain

[31] The relative skill of each interpolation method is summarized in Table 3, which shows domain-wide skill scores and a ranking for each variable. As scores for each test have different scales, we then average the ranks to produce an overall rating. On this basis, GK emerges as the best method for almost all climate variables except maximum temperature where TPS3D performs better. Although GK is the method with the best average rank, the differences between the top few methods are relatively

small. However, these domain-wide statistics may hide regional variations in relative skill that average-out when all stations are considered together. In the sections that follow, we therefore evaluate cross-validation skill spatially and through time.

4.2. Spatial Variations in Cross-Validation Skill

[32] We first evaluate whether there is any spatial pattern in the relative skill of the interpolation methods. Figure 1
shows, for each climate element, the best performing method on a station-by-station basis. For precipitation, GK is the best method at a large majority of stations, especially in Eastern Europe. In Western Europe, ADW2 and GK are roughly equal in the number of stations at which they are the best overall method; this is in agreement with the average statistics in Table 3. LK is the best method over the Netherlands, most likely due to the dense station network, and the lack of topographical complexity. In other areas LK suffers, because fewer data contribute to the estimation of the variogram.

[33] For all three temperature variables and pressure, NNI is the method that scores best at more stations than any other method, although it is not the method with the best average skill. The reason for this is that NNI has large errors at some other stations, which inflates the average skill scores. Several other methods have the best overall rank at a smaller fraction of stations: GK, LK, ADW and TPS3D. A large number of stations where TPS3D performs best are located around the Alps, reflecting the importance of using elevation as a co-predictor, but there is no other distinct location where a particular method is clearly the best.

Figure 2. CRE for precipitation for each station (left) and difference in CRE compared to the best overall method, GK. Histograms shows the frequency distribution of difference in CRE, with colors corresponding to scale bar below. Note that for CRE, a lower score indicates greater skill, so positive values at the right indicate GK has high skill.

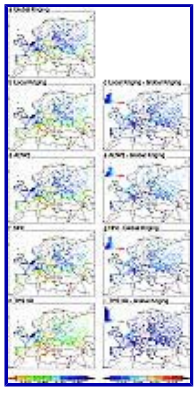

Figure 3. As in Figure 2, but showing results for CSI. Note that for CSI, a higher score indicates greater skill, so negative values at the right indicate GK has high skill.

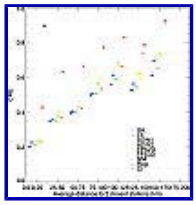

Figure 4. CRE as a function of average distance to the two closest stations for each interpolation method. For CI, CRE does not fit onto the scale.

[34] We next look at the difference in skill between interpolation methods, focusing on precipitation and mean temperature, but noting that similar results are obtained for all the temperature elements and MSLP. Figures 2 and 3
show the spatial distribution in CRE and CSI scores for precipitation for a subset of the interpolation methods, concentrating on those that perform best overall (Table 3). For each interpolation method the pattern of skill is similar. For precipitation, skill is best over the Netherlands and the UK and worst over Poland, northern Norway, Italy and Turkey. These last areas have a very sparse station network, which is likely the main reason for poorer skill. The reduced skill in areas with sparse station networks is confirmed in Figure 4, which shows a decline in CRE with distance from the nearest two stations; similar results (not shown) emerge for other skill scores.

[35] In the second column of Figures 2 and 3
the difference between the interpolation methods and GK are shown. In general, the differences are quite small, but it is clear that more stations have higher skill for GK than for any other interpolation method. The differences are generally larger for CSI than CRE, suggesting that skill is more variable for state than amount (but note that different skill scores have different scales). Other skill scores show similar patterns (not shown), with R, RMSE and high-CSI being similar to CRE and MAE, LEPS and PC being more similar to those of CSI. There are few obvious areas where other interpolation methods are doing better than GK. Over the Netherlands LK performs better, as already explained above, and ADW2 seems to perform better in the areas that have lowest CRE for GK, namely, areas with a low station density. The reason for this may be that GK, which has generally a higher search radius in these areas than ADW, uses stations for the interpolation with information that is not useful for the locations of the stations with low CRE.

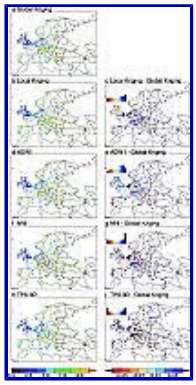

Figure 5. See Figure 2, but now RMSE for mean temperature.

[36] For the other variables the patterns of skill scores and differences from GK are similar, but differences are smaller than for precipitation. Figure 5 illustrates this using RMSE for mean temperature, and shows that GK performs better than the other interpolation methods for most stations. Interpolation skill is best over the UK and Germany and worst over the former Yugoslavia, Bulgaria, Romania, Northern Scandinavia and Northern Russia. Compared to the other interpolation methods, there are no clear areas where these methods perform better, although most seem to perform better for the west of the UK and Ireland and TPS3D performs better for the Alps. The behavior of the skill scores MAE and CRE are very much the same as for RMSE (not shown). However, GK does not perform noticeably better (or worse) than ADW, TPS3D or NNI when looking at the other skill scores (R, high-CSI, low-CSI and LEPS). Results for minimum and maximum temperature are very similar to the results for mean temperature. Differences between the variables are that skill is higher for mean and maximum temperature than for minimum temperature. For temperature there is also a decrease in skill with an increase in average distance to the two closest stations for all interpolation methods and skill scores (not shown); this effect is less obvious than for precipitation because the station distribution for temperature is more homogeneous over the domain. One interesting feature is that the decline in skill for TPS3D at large station distances is much less than for other methods, suggesting that inclusion of elevation as a co-predictor is important when interpolating in data-sparse areas.

[37] For MSLP, for which only few station data are available, the skill is highest in Germany and lowest in the Netherlands and at single stations in Turkey and Israel. GK is the best method for all skill scores. One clear difference with other methods is that ADW seems to perform better in the Alps for all skill scores.

4.3. Comparison Against Regional High-Resolution Gridded Data

[38] As noted earlier, we make use of high-quality regional gridded data to evaluate the ability of each interpolation method in producing accurate area-average grid estimates (as opposed to point-based station estimates evaluated using cross-validation). As with the station data we calculate skill scores between our gridded estimates, and the regional grids (Table 4). Overall, the skill scores are of similar magnitude for gridded data and station cross-validation. For some of the scores, such as CRE, R and CSI the gridded comparison yields higher skill than the cross-validation; for other skill scores, such as MAE and LEPS, the gridded comparison has lower skill than the cross-validation.

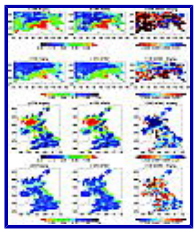

Figure 6. (left column) Spatial pattern of skill scores for GK and (middle column) ADW2 when validated against high-quality gridded data for the Alps and UK, and (right column) the difference in skill scores.

[39] Average skill is higher for each grid in the UK than the Alps, corresponding to the station cross-validation skill differences in the two regions. A key reason for this is the topographic complexity in the Alps, which makes interpolation more difficult [e.g., *Briggs and Cogley*, 1996]. In the Alps, ADW2 has higher average skill than GK (Table 4 and Figure 6); CRE for ADW2 is better in the north-western part of the Alps and CSI is better in the southern part of the Alps. This indicates that for many grids neither of the methods interpolates both state and magnitude of precipitation better than the other method. The skill in the North of Italy should be interpreted carefully; the Alps data set network is not particularly dense in this domain (indeed it is almost the same as our European data set). ADW2 may perform better in this area than GK because a similar ADW technique has also been used for development of the Alps data set. For the UK GK has higher skill than ADW2. ADW2 has a better skill in interpolating state than magnitude of precipitation compared to GK, since there are many more grids for which ADW2 performs better in Figure 6l (CSI) than in Figure 6i (CRE). GK mainly performs much better in the western part of Scotland. However, in the other mountainous areas of the UK, Snowdonia in north Wales, ADW2 has higher skill in CRE. Neither method appears to be better over all mountainous areas.

4.4. Seasonal Differences

[40] Tables 5 and 6 show the separate skill scores for winter (November-April) and summer (May-October). Comparing these tables shows that differences in skill between winter and summer are larger than differences in skill between the better performing interpolation methods. Skill for the interpolation of precipitation is on average better in winter than in summer, while for the temperature variables, the skill is better in summer than in winter. The signal for MSLP is not clear as some skill scores show better results in summer and others better in winter. The reason for better skill in the interpolation of precipitation in winter than in summer is that in winter most precipitation is frontal, whereas in summer precipitation is convective. For this reason, winter precipitation is more predictable than summer precipitation. Although there are large differences in skill in winter and summer, the rank order of the interpolation methods has not changed: in general, the same interpolation methods perform best in both winter and summer. The only exception is maximum temperature where GK is the best method in summer, and TPS3D is the best method in winter.

4.5. Skill Variations Over Time

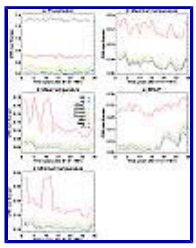

Figure 7. Variation in domain-wide CRE over time; CRE is calculated across all stations in the domain on a day-by-day basis, and then averaged by season for clarity in the display.

[41] As the quality of the station network varies over time, we also tested whether the absolute and relative skill of the interpolation methods changes over time. To do this we calculated skill scores over all available stations on a day-by-day basis. Figure 7 illustrates the results for CRE, showing the median skill score for each year. Regression and CI (precipitation only) have much higher CRE than the other interpolation methods, and these remain so over the entire analysis period. For mean and minimum temperature, the regression method shows much lower skill (peaks in CRE) at years 1, 5 and 12; these do not occur for other methods, indicating that they are not related to particular station values or changes in the station network. Additionally, since we show the median skill, they cannot be due to single outliers. Thus these peaks appear to reflect some instability in the regression interpolation method. Since regression is the worst method overall, we do not explore this further. Other, smaller periods with decreased skill are also visible for other variables, and tend to be common to all interpolation methods. For example the increase in CRE for precipitation at years 15 and 16 is caused by a short-term decrease in the density of the station network. The peak after year 18 for MSLP is due to the addition of around ten Belgian stations, which then do not have data after year 24. These stations cause deterioration in skill, probably because of an undetected data error common to all ten stations. The second peak in CRE for MSLP after year 26 is caused by reduced skill for some stations in the Alps. Again, this points toward problems with data quality at a group of local stations.

[42] Despite the variations in skill over time, the relative skill of the different methods remains relatively constant. Precipitation CRE for GK and ADW2 are nearly identical for the entire record. For 18 out of the 30 years ADW2 has slightly lower skill than GK, while for the other 12 years ADW2 skill is higher. For temperature the best performing method is either GK or TPS3D. For mean and minimum temperature GK is the best performing method for 28 and 27 years out of the 30 in total, respectively, while for maximum temperature TPS3D is the method with the best skill for 17 years. For MSLP GK is constantly the best method, until year 27. After that, in a three-year period with higher skill than before, LK and TPS2D are the better methods. Similar results emerge for the other skill scores (not shown).

4.6. Skill at Estimating Extremes

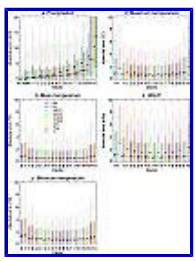

Figure 8. Absolute error in different deciles for each variable and interpolation method. The dot shows the median error in each decile; thick lines run from the 10th to the 90th percentile and the dashed lines from the 1st to the 99th percentile.

[43] So far, we have evaluated interpolation skill considering all station data at the same time, but we would like to be confident that the method that is best overall also performs best in estimating extreme events. To do this, we sort data from each station into deciles, and calculate the mean estimation error for each station decile. For precipitation, deciles are calculated for non-zero data (wet days), but we also show mean error for dry days. For precipitation, analysis is undertaken on the absolute amounts, while

for the other variables, data are first converted to anomalies from the mean. Results are summarized for all stations and interpolation methods in Figure 8.

[44] For all variables it can be seen that extremes have the highest absolute error; similar results are found for other skill scores (not shown). In addition, as the absolute error increases, differences between interpolation methods become exaggerated. As in previous sections, regression and CI (for precipitation) have by far the largest error. For the other interpolation methods the differences are much smaller. For all variables NNI has small median absolute error, as do GK and ADW. In fact, NNI has the lowest median absolute error for all percentiles of maximum temperature and for the majority of deciles for SLP. For precipitation LK has smaller median absolute error than GK for the 2nd to 5th deciles, while GK has the smallest error for the 6th to 9th deciles. When there is no rain, which is on approximately two thirds of the days, GK, ADW2 and CI all have a median absolute error of zero, indicating that most of the dry days are correctly estimated. The largest errors follow a similar pattern to the median errors, increasing toward higher precipitation and larger anomalies for the other variables. For precipitation, GK has the lowest range of absolute error at low precipitation amounts, while ADW2 has the lowest range at higher precipitation amounts. For temperature and MSLP GK generally has the lowest range of absolute error, while NNI tends to have a larger range, explaining why NNI is overall not the best method for the interpolation of temperature and MSLP, even though it is the best methods for many stations in Europe (Figure 1).

[45] Table 3
also provides insight into the relative performance of interpolation methods in relation to the extremes of the variable. CSI-low and CSI-high provide information on the proportion of time both observed and estimated values fall below/above the 5th/95th percentiles. GK is generally very good at interpolating these extreme values, while NNI in general performs well, because it smoothes less than, for example TPS.

5. Discussion and Conclusions

[46] We have described the comparison of nine different versions of six interpolation methods, regression, NNI, ADW (1 and 2), kriging (global and local), TPS (2D and 3D), and CI, for five climate variables (mean, minimum and maximum temperature, precipitation, and MSLP) at daily time steps over Europe. We first evaluated relative skill of each method through station cross-validation, looking at (1) average skill over the entire domain, (2) spatial patterns of skill at individual stations, (3) variations in skill over time as the station network varies, and (4) skill for different deciles in the frequency distribution at each station. We also compared, for precipitation, the candidate interpolation methods against existing high-resolution gridded data sets for the UK and the Alps, which are derived from a much denser network of stations. For all comparisons, we use a range of skill scores that evaluate different aspects of estimation skill, particularly for precipitation, where we need to evaluate skill in estimating state (wet/dry days) as well as amount.

[47] Apart from regression and CI, the differences between the interpolation methods for all skill scores and all variables are fairly small. In fact, the differences in skill between summer and winter for a single method are larger than the differences between the methods. In addition to seasonality, skill is also influenced by station density, with all interpolation methods performing better in areas with higher-density station networks. However, not all interpolation methods respond in the same way to changing station networks, for example, while GK and ADW2 have very similar skill when the station network is dense, ADW2 performs better when nearest stations are more distant. Skill in areas that are topographically complex tends to be poorer, as shown by the station cross-validation in the Alps, and the comparison to high-resolution gridded data over the Alps and UK; for the latter skill is lowest in mountainous areas of Scotland and Wales. In these areas, different interpolation methods perform slightly better, depending on the variable of interest and the station density. However, the potential improvements in interpolation accuracy that might arise from using different interpolation methods in different areas or for different variables do not seem large enough to warrant such a complex approach to generating a pan-European gridded data set.

[48] Overall, GK is the best ranked method for all climate variables except maximum temperature, where TPS3D has a marginally better overall skill (Table 3). NNI performs well at many stations for the temperature variables and SLP, but average skill for NNI is negatively affected by larger errors at stations where NNI performs relatively poorly. Many studies have found that universal kriging is one of the best interpolation methods for both mean precipitation and temperature [e.g., *Attorre et al.*, 2007], as well as daily climate variables [*Jarvis and Stuart*, 2001; *Stahl et al.*, 2006]. *Stahl et al.* [2006] compare 12 variations of regression-based, kriging and weighted average approaches for the interpolation of daily minimum and maximum temperature over British Columbia, Canada. They find that the GIDS method (a method based on multiple linear regression) performs best when a high station density network is available. However, they conclude that methods that compute local lapse rates from the control points, like GIDS, should not be applied in the absence of sufficient higher-elevation station data because these methods performed more poorly for the years for which there were a smaller number of higher-elevation stations. For the years with lower station density, their ordinary kriging method performed best.

[49] The skill (absolute or relative to other methods) of the interpolation is relatively constant in time For MSLP, skill decreases in the 1980s, but the reasons for this are the data quality issues at stations in Belgium and the Alps, which are being investigated further prior to construction of a final data set. There are also only a few areas where a specific interpolation method appears to have higher skill: in the Netherlands LK appears to be the best method for precipitation and in the Alps TPS3D appears better for temperature, but in these areas differences between these methods and GK remain small.

[50] Thus, while there is no interpolation method that stands out as superior to others by a large margin and several methods perform best when considering a specific criterion, climate variable or sub-domain, GK is by a small margin the best overall method. GK also has the advantage, along with LK and TPS of yielding interpolation error estimates (analogous to the confidence intervals for regression). Therefore global kriging was selected as the method to be used in the construction of the ENSEMBLES 0.2° longitude/latitude gridded daily climate data set, described in detail by *Haylock et al.* [2008].

Acknowledgments

[51] We would like to thank all institutes (see Appendix 1 of *Klok and Klein Tank* [2008]) that made meteorological station data available for the study. This study was funded by the EU project ENSEMBLES (WP 5.1, contract GOCE-CT-2004-50539). NH was also funded by the Dutch talentenbeurs and the Dutch Prins Bernhard Cultuurfonds beurs.

Citation: Hofstra, N., M. Haylock, M. New, P. Jones, and C. Frei (2008), Comparison of six methods for the interpolation of daily, European climate data, *J. Geophys. Res.*, *113*, D21110, doi:10.1029/2008JD010100.