Testing E-OBS European high-resolution gridded dataset of daily precipitation and surface temperature

Nynke Hofstra[1,2,*], Malcolm Haylock[3], Mark New[1], Phil D. Jones[3]

[1] School of Geography and the Environment, Oxford University Centre for the Environment, South Parks Road, Oxford, OX2 8JU, England

[2] Now at: Environmental Systems Analysis Group, Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands

[3] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, England

[*] Corresponding author: nynke.hofstra@wur.nl

Abstract

Gridded datasets derived through interpolation of station data have a number of potential inaccuracies and errors. These errors can be introduced either by the propagation of errors in the station data into derived gridded data or by limitations in the ability of the interpolation method to estimate grid values from the underlying station network. Recently, *Haylock et al* [2008] reported on the development of a new high-resolution gridded dataset of daily climate over Europe (termed E-OBS). E-OBS is based on the largest available pan-European dataset and the interpolation methods used were chosen after careful evaluation of a number of alternatives, yet the dataset will inevitably have errors and uncertainties. In this paper we assess the E-OBS dataset with respect to: 1) homogeneity of the gridded data; 2) evaluation of inaccuracies arising

from available network density, through comparison with existing datasets that have been developed with much denser station networks; and 3) the accuracy of the estimates of interpolation uncertainty that are provided as part of E-OBS.

We find many inhomogeneities in the gridded data that are primarily caused by inhomogeneities in the underlying station data. In the comparison of existing data with E-OBS we find that while correlations overall are high, relative differences in precipitation are large, and usually biased towards lower values in E-OBS. From the analysis of the interpolation uncertainties provided as part of E-OBS, we conclude that the interpolation standard deviation provided with the data significantly underestimates the true interpolation error when cross-validated using station data, and therefore will similarly underestimate the interpolation error in the gridded E-OBS data. While E-OBS represents a valuable new resource for climate research in Europe, users of the data need to be aware of the limitations in the dataset and use the data appropriately.

1. Introduction

Gridded climate data derived from meteorological station measurements underpin a wide range of applications and research in climate science, including evaluation of global and regional climate models, the construction of bias-corrected climate change scenarios and driving many applications in climate impacts assessments [*Haylock et al.*, 2008]. Increasingly, there has been a need for gridded data at higher spatial and temporal resolutions, as the focus of climate change research has shifted from global to regional and local scales. Recently, *Haylock et al.* [2008] described the development of the first high-resolution gridded dataset of daily climate over Europe (termed E-OBS), as part of the EU funded ENSEMBLES project. The dataset, comprising daily mean, minimum and maximum temperature and precipitation, was constructed through interpolation of the most complete collection of station data over wider Europe [*Klok and Klein Tank*, 2008]. The data are available on four different RCM grids (0.25 and 0.5 degree regular lat-lon and 0.22 and 0.44 degree rotated-pole) and cover the period 1950-2006. Additionally, estimates of interpolation uncertainties are included as part of the dataset [*Haylock et al.*, 2008].

Gridded datasets derived through interpolation of station data have a number of potential inaccuracies and errors. Errors in the underlying station data can be propagated into the gridded data; typical sources of error include incorrect station location information, individual erroneous values or non-climatic breaks (inhomogeneities) in the station time series. A second source of uncertainty relates to the ability of the interpolation method to estimate grid values from the underlying station network. In general, interpolation accuracy decreases as the network density decreases, is less accurate for variables with more variable spatial characteristics (e.g. precipitation) and degrades in areas of complex terrain (e.g. mountain areas). While E-OBS is based on the largest available pan-European dataset and the interpolation methods used were chosen after careful evaluation of a number of alternatives [*Hofstra et al.*, 2008], the dataset will inevitably have errors and uncertainties.

The aim of this paper is to assess the E-OBS dataset with respect to some of the potential errors that may be present. Users can then familiarise themselves with the strengths and weaknesses of the data and use them responsibly. We have chosen three features of E-OBS to analyse in this paper: 1) homogeneity of the gridded data; 2) inaccuracies due to the underlying station network density, though comparison with existing datasets that have been developed with much denser station networks; and 3) the accuracy of the estimates of interpolation uncertainty that are provided as part of E-OBS.

Long-term station data are often influenced by non-climatic factors, such as changes in station location or environment, instruments and observing practices. These so-called inhomogeneities can often lead to misinterpretations of the climate data analysed [*Peterson et al.*, 1998]. The station data used for E-OBS are not fully homogenised. Individual station series may have been homogenised by the original custodians of each series, but the series provided by partner organisations have been used directly, meaning potentially inhomogeneous stations may be contributing to the interpolated grids. As station density strongly influences the interpolation [*Hofstra et al.*, 2008], E-OBS was constructed using many *potentially* inhomogeneous stations, as their exclusion would degrade the station network density and hence accuracy of the interpolation. In addition, several studies explain that, for area averages of relatively large areas, inhomogeneities balance out during interpolation [*Dai et al.*, 1997; *New*, 1999; *Peterson et al.*, 1998]. However, that may not be the case for the E-OBS high-resolution grids. Therefore, the first out of three topics tested is the homogeneity of the dataset.

The second topic is a comparison with other gridded datasets that have been developed with much denser station networks. These datasets are available, in the case of precipitation, for long periods for the UK and the Alps and for the period October 1999 – December 2000 for Europe as a whole. For temperature, unfortunately, we have only been able to secure data for the UK. Datasets developed with denser station networks are assumed to be a better approximation of the true area-averages. So if the E-OBS gridded dataset produces grid area-averages that are close to those calculated from the higher quality grids, the E-OBS dataset can be deemed to be a reasonable representation of the true area-average gridded values.

Because of the inevitable interpolation uncertainties, the E-OBS dataset is provided with information on the interpolation uncertainty for each grid box and each day [*Haylock et al.*, 2008]. E-OBS interpolation uncertainty is derived by combining the Bayesian standard error estimates of the monthly climatology [*Hutchinson*, 1995] and the interpolation standard deviation for daily anomalies [*Yamamoto*, 2000] (see section 5 for more detail). Here we concentrate on the interpolation standard error estimates, and evaluate the accuracy of the estimates through cross-validation against station data. This represents the first evaluation of the *Yamamoto* [2000] standard error method, which has to date only been applied to geological data.

The remainder of the paper is structured as follows. Section 2 provides a more detailed description of the E-OBS dataset, including the underlying station data and the interpolation and gridding methodology. We then cover each of the three evaluations in turn: inhomogeneities (Section 3), comparison against regional gridded datasets based on denser station networks (Section 4) and evaluation of the interpolation standard error estimates (Section 5). We conclude with a summary of results and a discussion of the implications of our assessment for use of the E-OBS dataset.

2. The E-OBS dataset

The E-OBS gridded dataset is derived through interpolation of the ECA&D (European Climate Assessment and Data) station data described in *Klok and Klein Tank* [2008]. The station dataset comprises a network of 2316 stations, with the highest station density in Ireland, the Netherlands and Switzerland, and lowest density in Spain, Northern Africa, the Balkans and Northern Scandinavia. The number of stations used for the interpolation differs through time and by variable. The full period of record used for interpolation is 1950 – 2006 , but the period 1961 –

1990 has the highest density. At any particular time, there are more precipitation than temperature stations. Inhomogeneities in the station time-series have been flagged, but potentially inhomogeneous stations are used for the interpolation, for reasons noted above.

The E-OBS dataset is derived through a three stage process [*Haylock et al.*, 2008]. Monthly means (totals) of temperature (precipitation) are first interpolated to a 0.1 degree latitude by longitude grid using three-dimensional (latitude, longitude, elevation) thin plate splines. Daily anomalies, defined as the departure from the monthly mean (total) temperature (precipitation), are interpolated to the same 0.1 degree grid, and combined with the monthly mean grid. For temperature, daily anomalies are interpolated using kriging with elevation as an external drift factor. For precipitation indicator kriging is first used, where the state (wet/dry) of precipitation is first interpolated, after which the magnitude at 'wet' 0.1 degree grid points is interpolated using universal kriging. Finally, the 0.1 degree points are used to compute area-average values at the four E-OBS grid resolutions (0.25 and 0.5 degree regular latitude-longitude grid and 0.22 and 0.44 degree lat-long rotated-pole grids). In this paper, we use the 0.25 degree regular latitude-longitude grid for further evaluation, as results for the other grids are essentially the same.

Standard error estimates that accompany the gridded data are derived through combination of the individual standard error estimates for monthly and daily interpolations. Standard error for the monthly mean or total are the Bayesian standard error estimates, as available in the ANUSPLIN package used for the spline interpolation [*Hutchinson*, 1995; *Wahba*, 1983]. Error estimates for daily anomalies have been calculated using the method proposed by *Yamamoto* [2000] (see Section 5). Both standard error estimates are calculated at the 0.1 degree master grid. For temperature monthly and daily uncertainties are combined taking the square root of the sum of the squares of the two uncertainties. For precipitation the relative uncertainty of the daily total is the square root of the sum of the squares of the relative uncertainty of the monthly total and the relative uncertainty of the daily proportion of monthly total precipitation. Uncertainties at the 0.1 degree grid have been averaged over the target grids allowing for spatial autocorrelation. Details on the interpolation methods and how we implemented them as well as on the calculation of the uncertainties are available in *Haylock et al.* [2008].

3. Homogeneity assessment

3.1. Homogeneity testing

To analyse the influence of inhomogeneities in station data on gridded time-series and to inform the user about possible inhomogeneous areas within the dataset, we apply a homogeneity test to the gridded dataset and compare results to the same test for station data. Numerous tests could be used [e.g., *Peterson et al.*, 1998], but for this study we use the Wijngaard method [*Wijngaard et al.*, 2003], which is the same test that was applied to the ECA&D station data used to construct the E-OBS, where 39% of the precipitation and 25% of the temperature station series were found to be potentially homogeneous over the period 1961 – 2006 [*Klok and Klein Tank*, 2008].

The Wijngaard method is an absolute test, as it does not use a supposedly homogeneous reference series. This was appropriate for the version of the ECA&D dataset before the ENSEMBLES project started, because of its sparse network [*Wijngaard et al.*, 2003]. It comprises four homogeneity tests: the standard normal homogeneity test (SNHT) for a single break [*Alexandersson*, 1986], the Buishand range test [*Buishand*, 1981], the Pettitt test [*Pettitt*, 1979] and the Von Neumann test [*Von Neumann*, 1941]. These location-specific tests have different characteristics; for example, the SNHT test is more sensitive to inhomogeneities earlier or later in the time-series, whereas the Buishand and Pettitt tests work better for breaks near the middle of the series. If zero or one of the tests detects a break at the 1% significance level the time-series is classified 'useful'; if a break is detected by two tests the series is classified 'doubtful' and if three or four tests find a break, the series is classified 'suspect'.

For precipitation the annual wet day count is used for the analysis of breaks, as this statistic generally has lower variance than total precipitation, enabling a better signal to noise ratio for significance testing. For temperature, the annual mean diurnal temperature range (mDTR) and the annual mean of the absolute day-to-day differences of DTR (vDTR) are used for homogeneity detection. DTR is used in preference to mean, maximum or minimum temperature, as it has been shown that tests on DTR are more sensitive: breaks that are mainly radiation related have different effects on minimum and maximum temperature and are, therefore, only weakly apparent in these variables, but do appear clearly in DTR. As the homogeneity tests are applied to both mDTR and vDTR, a temperature station is classified according to the worst outcome for the two variables.

We apply the Wijngaard tests to both station and E-OBS gridded data and compare the results. We calculate the annual wet day count, mDTR and vDTR for each year if for each month no more than 20% of the data are missing. If less than 80% of the years in the period 1950-2006 are present, the homogeneity test for that station or grid box is not performed, although these stations may have been used for the interpolation. *Wijngaard et al.* [2003] concluded that a 1 mm threshold should be applied to define a wet day because otherwise too many breaks were detected, and we accordingly adopt this threshold.

3.2. Results and discussion

Figure 1 shows the stations and grid boxes that are potentially useful (green), doubtful (blue) or suspect (red), according the Wijngaard classification. For precipitation there are more many more useful stations and grid boxes than suspect ones. Suspect areas are mainly located in Northern Norway, Scotland, Italy, the Balkan, parts of Central Europe and in Northern Russia. For temperature most of Europe has a statistical significant inhomogeneity at some point in the gridded data, indicated by breaks in mDTR or vDTR (or both). However, if we only look at mDTR there are major differences (see Figure-S 1 in the supplementary material), with many more potential homogeneities in coastal areas, with remaining areas of central France, UK, Netherlands, parts of Spain and major parts of Ukraine, Northern Russia, Finland, southern Sweden, Czech Republic, Baltic States and Former Yugoslavia classified as useful in that case. That we find breaks in mDTR along the coast may be explained by a reduced variability in those areas due to the influence of the sea, making it easier to detect a break in mDTR. Inhomogeneities are much more widespread in vDTR with no clear difference between coastal and non-coastal areas.

Figure 1 also shows that the areas that have the most suspect stations often also have suspect grids, but sometimes even one suspect station may influence a whole area. An example of the latter is precipitation in northern Sweden where only one station is suspect, but has an influence over many grid boxes. Conversely, some stations have a smaller influence on the area, as, for example, in Russia where many stations are inhomogeneous, but only small areas are influenced. Many stations in this area have breaks in different years and these may be cancelled out in the gridded values. For temperature, inhomogeneous stations are present across the whole of Europe, which is reflected in the inhomogeneities of the gridded data.

In the case of precipitation many more areas of the grids are classified as potentially useful than for temperature (78% for the wet day count versus 46% for mDTR and 28% for vDTR for the grids, and 89% versus 49% and 56% for the stations, see Table 1), which is related to the fact that the homogeneity test is less sensitive for the wet day count. The percentage of stations that are qualified useful is higher in this study than in the study of *Klok and Klein Tank* [2008] (89% for the wet day count in this study vs. 39% in the *Klok and Klein Tank* study and 49% vs. 25% for temperature). The reason for this is most likely the time period used; we use the additional first 11 years of the data, in which fewer stations have full data coverage. When there are fewer stations available, also fewer breaks are detected in the data. mDTR has a much higher percentage of useful grids than vDTR, whereas vDTR has a higher percentage of useful stations than mDTR. This indicates that in the station breaks are more strongly manifested in the mean of the data, whereas in the grids breaks are more strongly manifested in the standard deviation. That may be due to the fact that the variability of the grid values are dependent on the station density of the network used for the interpolation and the distance to the grid centre [*Hofstra et al.*, 2009]. A station network that does not have a constant density in time may introduce inhomogeneities.

We also assessed the distribution of breaks in time and compare these between gridded and station data (Figure 2). As expected, the SNHT detects more inhomogeneities near the beginning and end of the period than the Buishand and Pettitt tests. SNHT also detects more breaks for any one variable than the other tests (Table 1). For wet day count the inhomogeneity in 1965 detected in the station data by the Pettitt test is also visible in the gridded data. Breaks in the 1975-1985 period in the station data are mainly reflected in the gridded data close to 1980. For mDTR the breaks in station and gridded data do not show a specific pattern. However, where for vDTR the largest inhomogeneities in the station data are found around 1970, the largest breaks in the gridded data are found in the early 1990s. The latter breaks may be due to a declining station density around this time. We investigated whether inhomogeneities could be determined on a decadal basis, by analysing each of the six decades separately, but the Wijngaard method is not sensitive enough to find any inhomogeneities in these shorter periods at the 0.01 significance level.

We also divided the calculated potential breaks for all three methods of the 57 year period into six decadal groups and assess the inhomogeneities spatially (see Figures S2-S5 in supplementary material). We can conclude, for example for precipitation, that most Italian and former Yugoslavian stations around the Adriatic Sea with a break have this break in the period 1980-1990 for all three tests; these breaks are also propagated through into the gridded data. For precipitation, for all three tests in general, the timing of the breaks in the gridded and station data compares quite well. For temperature, the agreement in timing of breaks between the station and gridded data is smaller. For example, for vDTR a large part of Russia and the Ukraine has the largest significant break between 1990 and 2000 for all three tests, whereas most stations in this area suggest the largest break exists between 1960 and 1980. This indicates that there may be multiple breaks in the station time-series of which one becomes more important in the gridded data.

The inhomogeneities within the gridded data are important to keep in mind during any use of the dataset. For example, when studying trends in the data, the results within the areas that are suspect may not be meaningful. For those who require more detail on the inhomogeneities in the gridded data, we have prepared a file that includes, for precipitation and temperature, the potential classification of homogeneity of each 0.25 degree grid box (useful, doubtful, suspect) and, for each of the four homogeneity tests, whether a statistical significant inhomogeneity has been detected and if so the year of the largest break. The file can be downloaded from the E-OBS download site (http://eca.knmi.nl/download/ensembles/ensembles.php).

4. Comparison with existing datasets

4.1. Existing datasets

In the second test of the dataset we compare E-OBS to existing datasets developed with much denser station networks. Since station density is a very important factor in the interpolation and the interpolation errors are smaller in areas with a dense station network [*Hofstra et al.*, 2008], these existing datasets are deemed close to the 'true' areal average, and provide a useful reference against which to judge the E-OBS dataset. The three existing datasets used are the UK, Alps and ELDAS datasets. ELDAS and the Alps datasets only comprise precipitation data. The UK dataset contains all four variables. We were unable to find or not allowed access to additional datasets in other regions.

4.1.1. UK

The UK dataset, supplied by the UK Met Office, comprises a 5x5 km equal-area grid, covering the period 1958 – 2002 for precipitation, 1995 – 2002 for minimum and maximum temperature and 1995 – 2006 for mean temperature [*Perry and Hollis*, 2005]. This dataset is compiled from a station network of 4400 stations for precipitation and 540 stations for temperature using multiple regression with geographic factors as the independent variables, followed by inverse distance weighting (IDW) of the residuals. In comparison, the ECA&D station network had 138 stations within this area, of which most had 70 - 85% of the data available for all variables. To allow comparison with the E-OBS interpolations all grid-points within each 0.25 degree grid used for the interpolation have been averaged. We also compare this dataset to ELDAS (see Section 4.1.3), for which a 1 degree grid is used.

4.1.2. Alps

The Alps dataset, comprising precipitation only, is an updated version of the climatology and daily data described by *Frei and Schär [1998]* and *Schwarb* [2001], described in more detail by *Hofstra et al.* [2008]. The data are available on a 0.25 by 0.1667 degree grid and cover the period 1966 – 1999. For the period 1966 – 1970 there are no data available over Austria and after 1990 there are data quality issues with many of the Italian stations, so in our comparison, we use the period 1966-1990, except for Austria, where the period 1971 – 1990 has been used. The dataset is constructed through addition of daily anomalies to the long term climatological mean. Anomalies were interpolated from station data using a modified version of the Shepard algorithm [an ADW technique, *Frei and Schär*, 1998; *Shepard*, 1984] and 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 dataset is based on over 6500 station records. In comparison, the E-OBS station network had 341 stations available within this area, with majority having over 70% data presence. To allow comparison with E-OBS on a common grid, both datasets have been averaged to a 0.25 x 0.25 degree grid.

4.1.3. ELDAS

The ELDAS daily precipitation dataset was developed by *Rubel et al.* [2004] for the Development of a European Land Data Assimilation System to predict Floods and Droughts (ELDAS) project. It covers Central and Northern Europe at 0.2 degree latitude by longitude and covers the relatively short period of October 1999 to December 2000. Some 21,600 stations were used for the interpolation, compared to 2000 for E-OBS over the ELDAS domain. Station density is reasonably homogeneous, but areas such as Portugal, Belgium, Italy, the Balkan, Czech Republic, the Baltic states and Scandinavia have a lower density than Spain, France, the Netherlands, the UK, Denmark, Germany, Poland, Switzerland and Austria. Interpolation was done via the Precipitation Correction and Analysis method [*Rubel and Hantel*, 2001]; this comprises a dynamical bias correction combined with an ordinary block kriging algorithm. To enable comparison, we averaged ELDAS and E-OBS to a common 1 degree latitude by longitude grid.

4.2. Comparison

We compare E-OBS to the high-quality grids using five skill scores for temperature and six for precipitation. We calculate the skill scores for all data together to obtain overall scores, and also on a grid-point basis to explore the spatial patterns in difference between grids. We use the mean absolute error (MAE), root mean squared error (RMSE), compound relative error (CRE) and Pearson correlation (R) to assess temperature and the precipitation amount. The Critical Success Index (CSI) and Percent Correct (PC) are used to study precipitation state (wet or dry, where a wet day is defined as having precipitation ≥ 0.5 mm). The skill scores are described in detail elsewhere [*Hofstra et al.*, 2008], but we include an explanation of each score in the supplementary material. For precipitation we also divide the MAE and RMSE by the mean precipitation for the grids in order to remove the influence of the amount of precipitation on these two skill scores in each grid.

We note that the high-quality data are not true areal averages. However, given they are based on order of magnitude denser networks than E-OBS, we expect them to be subject to smaller interpolation errors. Thus we can only quantify differences between the datasets, which provide a qualitative indication of potential errors in E-OBS, but should not be interpreted as errors of the dataset.

4.3. Results and discussion

Table 2 provides an overview of the results of the skill scores, calculated 'globally' for each grid pairing, as well as for each standard season. At first sight, the datasets compare very well: correlations, CSIs and PCs are high (for example, the global correlation coefficient for temperature is approximately 0.99 and for precipitation 0.85-0.92), the CREs are small and RMSEs are fairly small (for example, CRE is 0.02-0.04 and 0.18-0.36 for temperature and precipitation). However the mean differences between datasets are quite large. RMSE is 0.7-0.9 for temperature and 2.2-2.4 for precipitation, apart from the Alps where it is larger, at 5.8. MAE shows similar, but smaller differences. For precipitation, the relative RMSE varies between 0.73 (UK) to 1.3 over the Alps. Relative difference between E-OBS precipitation and the other datasets are smaller in winter (UK and ALPS) and autumn (ELDAS). The main reason for larger differences between the datasets in summer is that in summer precipitation is mainly convective rather than frontal. During this season the correlation between stations is lower than in the other seasons. Interpolation with a larger station density will then produce better areal averages than interpolation using a less dense network. For mean and minimum temperature the datasets are closer to each other in spring, whereas they compare better in winter for maximum temperature.

Figure 3 presents the results for precipitation spatially. E-OBS compares best to the UK dataset, as does the ELDAS dataset, suggesting that over the UK E-OBS is fairly reliable. The differences are generally larger over the West of Scotland, where topography is an important contributing factor to spatial variability in rainfall. E-OBS does not agree as well with the Alps dataset, where the topographic complexity means that the sparse E-OBS network does not result in the same gridded data as the denser Alps network; although absolute errors are large because precipitation is on average higher in the Alps, relative errors are also larger than in the UK.

Similarly, E-OBS compares poorly to ELDAS over Norway, due to the greater station density for the ELDAS dataset in this topographically complex area. Finally, the E-OBS precipitation dataset has virtually no stations available in northern Africa, which causes the poor agreement in this area. Figure 4 shows the spatial pattern of skill for temperature over the UK. In general, the agreement is good for all three temperature elements. Differences are greatest over Scotland compared to the rest of the UK. That may be a result of the higher station density of the UK network, which may have had more station data available at higher elevations in Scotland. Differences in agreement between the grids are generally larger than differences between the four seasons.

We also evaluate whether E-OBS shows a bias compared to the high density datasets, by counting the frequency of days where E-OBS is more than ± 0.1 standard deviations from the high density dataset (Figure 5). For precipitation, E-OBS shows a negative bias at nearly all grid boxes relative to the Alps and ELDAS datasets. Compared the ELDAS dataset, E-OBS is positively biased over parts of Norway and at scattered locations elsewhere in Europe. Over the UK, E-OBS rainfall tends to be negatively biased in areas of higher rainfall in the west, apart from Northern Ireland where there is a positive bias (and also compared to ELDAS). For temperature there are areas with a positive (too warm) and a negative (too cold) bias. One striking feature is that areas such as Devon/Cornwall and Southern Wales, that are too warm for minimum temperature, are often too cold for maximum temperature. The bias for temperature is not consistent over the whole of the UK.

In Figure 6 we assess the difference between E-OBS and the high density datasets across the distribution of precipitation amount and temperature. For this we calculate for each grid deciles of temperature and precipitation (for all wet days). We then calculate for each day and each grid the absolute difference between the E-OBS and the other datasets and plot the median, 5th, 25th, 75th and 95th percentiles of these differences in each decile (Figure 6). While precipitation is biased towards smaller values in all deciles of the dataset, the bias is larger for more extreme precipitation. In the comparison of the 10th decile for the Alps the error between the two datasets can be as high as 16 mm, which is the median of the error when E-OBS is compared to the Alps dataset (see median of 9-10th decile of E-OBS versus Alps comparison in Figure 6). The reason for this relates to the much higher station density in the other datasets. For E-OBS, interpolation typically occurs from more distant stations compared to the high density datasets; as extreme precipitation events are usually more localised, they will be over-smoothed if a sparse network is used. For temperature, differences in error are similar for all deciles, with an average of around 0.5 °C. The errors are slightly larger in the 1st decile for minimum temperature and the 10th decile for maximum temperature, which means that there are slightly larger errors in the extremes, but overall extreme temperature events will be quite well represented [see also the discussion of extremes in *Haylock et al.*, 2008].

We can conclude that the E-OBS shows quite large differences to the existing datasets based on higher density station network. While correlations overall, and on a grid-by-grid basis, are high, relative differences in precipitation are large, and usually biased towards an underestimation. For temperature (UK only), mean absolute differences are at least 0.5 °C. The fact that the ELDAS precipitation dataset shows a much better spatial match to the UK dataset than E-OBS underlines the fact that E-OBS is fundamentally limited by its underlying station network. As the E-OBS network density over the UK is above average compared to density over the rest of Europe, we can conclude that this issue is likely to be pervasive across much of the E-OBS domain. Assessment of the agreement with existing datasets for all deciles of precipitation and temperature shows that the errors are larger in the extremes than in the more average amounts of precipitation or temperature. There seem to be significant problems with the underestimation of precipitation extremes. Comparability is much higher for temperature than for precipitation, due to the fact that temperature is a continuous variable as opposed to precipitation.

5. Uncertainty assessment

5.1. Calculation of uncertainties

Brohan et al. [2006] give an overview of all sources of all known and calculable uncertainty in their HadCRUT3 gridded global monthly temperature dataset. Three groups of uncertainties have been identified: 1) station error, 2) sampling error and 3) bias error. Station error includes errors made during thermometer reading, possible adjustment of homogeneities, calculation of the station normal, and processing of raw data. The sampling error is the difference between the 'true' spatial average and the interpolated estimate. It depends on, amongst others, the number of stations in the grid box, the distribution of those stations and on the variability of the climate in the grid box. The gridding method used by *Brohan et al.* [2006] is a simple area average of the stations within a grid, which is different from the kriging method that we use, but the sampling error of our gridding method will depend on the same factors. Two sources of bias error are summarised by *Folland et al.* [2001]: urbanization effects [*Jones et al.*, 1990] and thermometer exposure changes [*Parker,* 1994]. For precipitation a similar list of sources of uncertainty can be made. Here we focus on sampling error as it is expected to be the largest contributor to overall error. The objective here is to evaluate the accuracy of the estimates of interpolation sampling error for daily anomalies used in E-OBS. As explained in the introduction, these daily errors are estimated using the method proposed by *Yamamoto* [2000].

Yamamoto [2000] estimates the so-called 'interpolation standard deviation' at each grid point as the weighted average of the squared differences between station and interpolated values as follows:

$$s_0 = \sqrt{\sum_{i-1}^{n} \lambda_i \left[z(x_i) - z*(x_0) \right]^2} \qquad [1]$$

where x_i (i =1,n) are the locations of the stations used for the interpolation and λ_i are the weights used in the kriging interpolation and z are the observed values at the i stations used for the interpolation (x_i) and z^* is the interpolated value at the location for the interpolation (x_0).

Yamamoto [2000] compared his interpolation standard deviation to the kriging standard deviation and cross validation error. The kriging standard deviation is a standard by-product of kriging and used widely as a measure of reliability of the kriging procedure. The interpolation standard deviation has much larger correlation with cross-validation error than with the kriging standard deviation. The reason for that is that the kriging standard deviation is not a true estimate of uncertainty [*Journel and Rossi*, 1989; *Monteira da Rocha and Yamamoto*, 2000], as it cannot properly measure local data dispersion [*Yamamoto*, 2000].

As we do not have the true grid values for evaluation, we adopt station cross-validation to test the accuracy of the *Yamamoto* [2000] interpolation standard deviation. We estimate the daily anomaly at each station in the ECA&D dataset used to construct E-OBS, using the same interpolation approach used for E-OBS gridded data. Interpolation standard deviation is calculated using equation [1] above and cross-validation error as the absolute difference between the interpolated station value and the observed value:

$$cve_0 = \left| z^*(x_0) - z(x_0) \right| \quad [2]$$

We next transform the interpolation standard deviations into 95% confidence intervals by multiplication with 1.96 (assuming a normal distribution) and addition to and subtraction from the interpolated daily values for each station. We then count the number of times the observed station value falls within the 95% confidence interval for the interpolated value, with the expectation that if the confidence interval is an accurate estimate of interpolation uncertainty we would expect the station value to fall outside the confidence interval approximately 5% of the time.

5.2. Results and discussion

We first compare the cross-validation error (CVE) and interpolation standard deviation (ISD) through scatter plots. Results are similar for all temperature variables, so we only show figures for precipitation and minimum temperature.

Correlation between the CVE and ISD for both temperature and precipitation is positive (Figure 7). The relationship between CVE and ISD is stronger for precipitation (r=0.57) than minimum temperature (r=0.33), which provides confidence that the spatial distribution of ISD will reflect the spatial variability in interpolation error. The relationship is also closer to one-to-one for precipitation, whereas for temperature, ISD tends to be too large at smaller CVE and vice versa.

However, a better test of the accuracy of the ISD is the count of the percentage of station values falling outside the interpolation 95% confidence interval derived from the ISD (Figure 8). For precipitation, the upper 95% limit is mostly exceeded between 5-10% of the time, while values fall below the lower limit 10-25% of the time, indicating that while the upper limit is a reasonable estimate, the lower limit is poorly defined, and that precipitation is frequently significantly underestimated. For temperature, there are roughly equal numbers of values falling above and below the 95% confidence interval, but as with precipitation, the number exceeds that expected. Most stations have at least 10% of data falling outside the confidence interval, with many stations having more than 25% of values outside the interval. There is also a clear north-south gradient in the percentage of the precipitation values falling outside the confidence limits, with the CI underestimation being much larger in the north. The main reason for this is the fact that there are fewer rain days in the south of Europe, compared to the north. The error is smaller when no or little precipitation is observed, compared to a situation when a lot of precipitation is observed.

From this analysis, we can conclude that the interpolation standard deviation provided with the data is a strong underestimation of the actual interpolation error and should be used with care. Moreover, it has to be taken into account, that the confidence intervals available with the gridded data only include interpolation sampling error and no station and bias errors.

6. Summary and Conclusions

We have analysed the new E-OBS European high-resolution gridded dataset of daily minimum, maximum and mean temperature and precipitation in three ways. First, we assessed the homogeneity of the gridded data and related this to the homogeneity of the station data. Secondly, we compared the dataset to existing gridded datasets developed with denser station networks. And finally, we evaluated the accuracy of the interpolation standard deviation, a measure of interpolation error that is provided with the dataset. While the three issues we assess do not give a complete overview of the reliability of the dataset, they do provide important additional information for users of the dataset.

The results of the Wijngaard [2003] homogeneity tests show that there are many *potential* inhomogeneities present in the gridded dataset. There are more statistically significant breaks present in temperature than precipitation data, and within the temperature data, there are more breaks for vDTR than mDTR variables. Inhomogeneities in the gridded data are often related to inhomogeneities in the stations contributing to the value of the grid. However, this relation is not the same for all areas. Sometimes an area is inhomogeneous even if there is only one inhomogeneous station in the area (e.g. for precipitation in northern Sweden) and in other occasions many stations are inhomogeneous, but the area is not effected (e.g. for temperature in south-eastern France). The year of the break of inhomogeneous grids generally corresponds to the year of the break of stations in the surrounding area, although the correspondence is better for precipitation than for temperature. We provide a data file that contains, for temperature and precipitation, information on the grid boxes where the data are potentially inhomogeneous. This information will be critical when, for example, performing analyses of trends in extremes using E-OBS. For a future update of the E-OBS dataset we recommend that the issue of inhomogeneities is studied thoroughly. A balance will have to be found between the loss of station data and the introduction of inhomogeneities and homogenisation of the station data should be considered.

When compared to existing high-resolution regional gridded data for the UK, ALPS and Europe (ELDAS) that are based on much denser station networks, E-OBS shows an excellent correlation. However, mean absolute errors are significant, in the order 0.5 °C for temperature and greater than 100% for precipitation. For both variables and all skill scores the datasets compare worse in areas with more relief. For precipitation agreement is in general better in winter, whereas for temperature agreement is mainly best in spring. In the case of precipitation, E-OBS also shows a negative bias, indicating that E-OBS tends to be over-smoothed relative to the high-density datasets. For temperature, E-OBS shows a small positive bias over quite large areas, but some scattered areas have a stronger negative bias. Moreover, the E-OBS dataset compares better to the mean of the variables of the existing datasets than to the extremes, although differences are much larger for precipitation than for temperature. Consequently, the dataset should be used with caution in comparison to RCM outputs, especially with respect to evaluation of RCM precipitation extremes.

The uncertainty estimates available with the data only represent sampling, or interpolation, errors. These are calculated by combining errors from both parts of the interpolation process, namely interpolation of the monthly mean (temperature) or totals (precipitation) using thin plate smoothing splines and the interpolation of daily anomalies using versions of kriging (see Section 2). We evaluated the daily interpolation error estimates, estimated using *Yamamoto*'s [2000] interpolation standard deviation approach. A comparison of these errors with cross-validation errors shows that for most of Europe cross-validation error is positively correlated with interpolation standard deviation. However, the frequency with which the 95% interpolation confidence interval is exceeded is much larger than expected, indicating that the interpolation standard deviation significantly underestimates the actual interpolation error. The 95% confidence limits are on average exceeded 25% and sometimes even over 50% of the time. In a future update of the data we recommend that ensemble stochastic simulations, i.e. a set of interpolated realisations should be considered for the estimation of uncertainties. These have also been mentioned in *Haylock et al.* [2008] but have not been implemented due to time constraints. *Bellerby and Sun* [2005] and *Teo and Grimes* [2007] suggest short-cuts that should reduce the computing time required.

The E-OBS dataset is the first publically available dataset that covers the whole of Europe at a very high spatial resolution for daily data. However, as this study reveals, there are some potentially important limitations to the data. Inhomogeneities are present within the data, the data show quite large absolute and relative differences and biases to existing datasets that have been developed with very dense station networks, and the standard errors delivered with the data appear to significantly underestimate the true interpolation error. This will have to be taken into account when the data are used, e.g. for the evaluation of RCM outputs. Trends analysis may also be affected by potential inhomogeneities in the data. In addition, the underestimation of extremes within the data may, for instance, influence future predictions using RCM outputs regarding flooding. Moreover, when using the standard errors that have been supplied with the data it has to be taken into account that these errors only include interpolation sampling errors and that they are an underestimation of the true error.

The E-OBS data will often be the only available dataset for studies of e.g. the comparison of RCM outputs for the whole of Europe. With the collation of more data and hence better availability, reconsideration of how to deal with inhomogeneities in station data and how to improve the uncertainty estimates the data will improve in the future. However, users of the data should take notice of the weaknesses mentioned in this paper and use the data appropriately.

Acknowledgements

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 is also funded by the Dutch Prins Bernhard Cultuurfondsbeurs and the Dutch talentenbeurs.

References

Alexandersson, H. (1986), A homogeneity test applied to precipitation data, *Journal of Climatology*, *6*, 661-675.

Bellerby, T. J., and J. Sun (2005), Probabilistic and ensemble representations of the uncertainty in an IR/Microwave satellite precipitation product, *Journal of Applied Meteorology*, *6*, 1032-1044.

Brohan, P., et al. (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *Journal of Geophysical Research*, *111*, D12106; 12101-12121.

Buishand, T. A. (1981), The analysis of homogeneity of long-term rainfall records in the Netherlands, KNMI Scientific Report WR 81-7, De Bilt, the Netherlands.

Dai, A., et al. (1997), Surface observed global land precipitation variations during 1900-88, *Journal of Climate*, *10*, 2943 - 2962.

Daly, C., et al. (2002), A knowledge-based approach to the statistical mapping of climate, *Climate Research*, *22*, 99-113.

Folland, C. K., et al. (2001), Global temperature change and its uncertainties since 1861, *Geophysical Research Letters*, *28*(13), 2621-2624.

Frei, C., and C. Schär (1998), A precipitation climatology of the alps from high-resolution rain-gauge observations, *International Journal of Climatology*, *18*, 873-900.

Haylock, M., et al. (2008), A European daily high-resolution gridded dataset of surface temperature, precipitation and sea-level pressure, *Accepted by Journal of Geophysical Research*.

Hofstra, N., et al. (2008), The comparison of six methods for the interpolation of daily, European climate data, *Accepted by Journal of Geophysical Research*.

Hofstra, N., et al. (2009), The influence of interpolation and station network density on the distribution and extreme trends of climate variables in gridded data, *Submitted to Journal of Climate*.

Hutchinson, M. F. (1995), Interpolating mean rainfall using thin plate smoothing splines, *International Journal of Geographical Information Systems*, *9*(4), 385-403.

Jones, P. D., et al. (1990), Assessment of urbanization effects in time series of surface air temperature of land, *Nature*, *347*, 169-172.

Journel, A. G., and M. E. Rossi (1989), When do we need a trend model in kriging?, *Mathematical Geology*, *21*, 715-739.

Klok, L., and A. M. G. Klein Tank (2008), Updated and extended European dataset of daily observations, *Accepted by International Journal of Climatology*.

Monteira da Rocha, M., and J. K. Yamamoto (2000), Comparison between kriging variance and interpolation variance as uncertainty measurements in the Capanema iron mine, state of Minas Gerais - Brazil, *Natural Resources Research*, *9*, 223-235.

New, M. (1999), Uncertainty in representing observed climate, in *Representing uncertainty in climate change scenarios and impact studies*, edited by T. Carter, et al., pp. 59-66, Climate Research Unit, Norwich.

Parker, D. E. (1994), Effects of changing exposure of thermometers at land stations, *International Journal of Climatology*, *14*, 1-31.

Perry, M., and D. Hollis (2005), The generation of monthly gridded datasets for a range of climate variables over the UK, *International Journal of Climatology*, *25*, 1041-1054.

Peterson, T. C., et al. (1998), Homogeneity adjustments of in situ atmospheric climate data: a review, *International Journal of Climatology*, *18*, 1493-1517.

Pettitt, A. N. (1979), A non-parametric approach to the change-point detection, *Applied Statistics*, *28*, 126-135.

Rubel, F., and M. Hantel (2001), BALTEX 1/6-degree daily precipitation climatology 1996-1998, *Meteorology and Atmospheric Physics*, *77*, 155-166.

Rubel, F., et al. (2004), Daily and 3-hourly Quantitative Precipitation Estimates for ELDAS, edited, p. 32, Biometeorology Group, Univ. Vet. Med., Vienna.

Schwarb, M. C., et al. (2001), *Mean annual and seasonal precipitation in the European Alps 1971– 1990*, plates 2.6, 2.7 pp., Landeshydr. und Geol., Bern.

Shepard, D. S. (1984), *Computer mapping: The SYMAP interpolation algorithm*, 133-145 pp., Springer, New York.

Teo, C.-K., and D. I. F. Grimes (2007), Stochastic modelling of rainfall from satellite data, *Journal of Hydrology*, *346*, 33-50.

Von Neumann, J. (1941), Distribution of the ratio of the mean square successive difference to the variance, *Annals of Mathematical Statistics*, *12*, 367-395.

Wahba, G. (1983), Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline, *Journal of the Royal Statistical Society. Series B (Methodological)*, *45*(1), 133-150.

Wijngaard, J. B., et al. (2003), Homogeneity of 20th century European daily temperature and precipitation series, *International Journal of Climatology*, *23*, 679-692.

Yamamoto, J. K. (2000), An alternative measure of the reliability of ordinary kriging estimates, *Mathematical Geology*, *32*(4), 489-509.

Table 1. The fraction of stations or grids that are useful, doubtful or suspect and the inhomogeneous fraction for each statistical test

		# stations or grids	Overall Fraction			Fraction with Breaks			
			Useful	Doubtful	Suspect	SNHT	Buishand	Pettitt	Von Neumann
Wet day fraction	Stations	836	0.892	0.044	0.064	0.123	0.072	0.114	0.087
	Grids	22176	0.781	0.078	0.141	0.219	0.164	0.216	0.166
mDTR	Stations	472	0.492	0.114	0.394	0.477	0.422	0.432	0.468
	Grids	21970	0.464	0.099	0.437	0.515	0.470	0.460	0.485
vDTR	Stations	472	0.555	0.097	0.348	0.434	0.388	0.400	0.381
	Grids	21970	0.275	0.113	0.612	0.738	0.630	0.580	0.697

Table 2. Skill scores for the comparison of the E-OBS gridded dataset with the UK, Alps, and ELDAS gridded datasets for the four variables minimum, maximum and mean temperature and precipitation. Skill scores have been calculated for each grid point and are then averaged.

Annual

		R	MAE	MAE/ mean	RMSE	RMSE/ mean	CRE	CSI	PC
UK	Minimum temperature	0,984	0,687	n/a	0,895	n/a	0,041	n/a	n/a
	Maximum temperature	0,991	0,597	n/a	0,780	n/a	0,024	n/a	n/a
	Mean temperature	0,991	0,517	n/a	0,695	n/a	0,023	n/a	n/a
	Precipitation	0,916	1,081	0,355	2,170	0,729	0,182	0,836	0,909
Alps	Precipitation	0,880	2,253	0,514	5,766	1,325	0,357	0,769	0,897
Eldas	Precipitation	0,846	1,159	0,457	2,419	1,009	0,316	0,744	0,874

Winter

		R	MAE	MAE/ mean	RMSE	RMSE/ mean	CRE	CSI	PC
UK	Minimum temperature	0,971	0,700	n/a	0,918	n/a	0,082	n/a	n/a
	Maximum temperature	0,977	0,507	n/a	0,680	n/a	0,056	n/a	n/a
	Mean temperature	0,974	0,533	n/a	0,718	n/a	0,068	n/a	n/a
	Precipitation	0,925	1,187	0,331	2,227	0,627	0,176	0,856	0,914
Alps	Precipitation	0,894	2,013	0,505	5,031	1,274	0,346	0,784	0,906
Eldas	Precipitation	0,848	1,256	0,458	2,360	0,926	0,373	0,759	0,869

Spring

		R	MAE	MAE/ mean	RMSE	RMSE/ mean	CRE	CSI	PC
UK	Minimum temperature	0,973	0,663	n/a	0,860	n/a	0,069	n/a	n/a
	Maximum temperature	0,981	0,640	n/a	0,822	n/a	0,051	n/a	n/a

	Mean temperature	0,984	0,491	n/a	0,631	n/a	0,039	n/a	n/a
	Precipitation	0,916	0,893	0,359	1,803	0,730	0,181	0,828	0,908
Alps	Precipitation	0,881	2,237	0,514	5,345	1,231	0,365	0,775	0,888
Eldas	Precipitation	0,853	1,039	0,465	2,103	0,992	0,338	0,742	0,875

Summer

		R	MAE	MAE/ mean	RMSE	RMSE/ mean	CRE	CSI	PC
UK	Minimum temperature	0,955	0,668	n/a	0,866	n/a	0,116	n/a	n/a
	Maximum temperature	0,970	0,709	n/a	0,896	n/a	0,087	n/a	n/a
	Mean temperature	0,965	0,520	n/a	0,700	n/a	0,082	n/a	n/a
	Precipitation	0,898	1,004	0,402	2,136	0,874	0,207	0,807	0,903
Alps	Precipitation	0,852	2,531	0,546	6,088	1,385	0,392	0,732	0,878
Eldas	Precipitation	0,826	1,026	0,514	2,003	1,334	0,577	0,690	0,870

Autumn

		R	MAE	MAE/ mean	RMSE	RMSE/ mean	CRE	CSI	PC
UK	Minimum temperature	0,976	0,720	n/a	0,928	n/a	0,067	n/a	n/a
	Maximum temperature	0,987	0,518	n/a	0,667	n/a	0,035	n/a	n/a
	Mean temperature	0,983	0,526	n/a	0,709	n/a	0,042	n/a	n/a
	Precipitation	0,921	1,243	0,341	2,408	0,681	0,173	0,849	0,912
Alps	Precipitation	0,899	2,228	0,495	6,196	1,368	0,326	0,783	0,914
Eldas	Precipitation	0,863	1,226	0,431	2,511	0,911	0,306	0,765	0,879

Figure 1. Overall homogeneity, according to the Wijngard test, of the station network (top) and the gridded data (bottom) for precipitation (left) and temperature (right). For temperature mDTR and vDTR are combined, with the most negative outcome for the two variables used.

Figure 2. The fraction of stations and grid points with a statistically significant (0.01) inhomogeneity in each year of the dataset. Inhomogeneities are calculated for the full 1950-2006 period.

Figure 3. A spatial overview of the skill scores R (-), MAE (mm), RMSE (mm), CRE (-) and CSI for precipitation for the comparison of the E-OBS dataset with the datasets of the UK (top row), Alps (2^{nd} row) and ELDAS (3^{rd} row) and the UK versus ELDAS (bottom row). MAE / mean precipitation (-) and RMSE / mean precipitation (-) are added to remove the influence of the average amount of precipitation in a grid cell on the skill score.

Figure 4. As Figure **3**, but for the skill scores R (-), MAE (°C), RMSE (°C) and CRE (-) for minimum (top), maximum (middle) and mean (bottom) temperature for the comparison with the UK dataset.

Figure 5. Spatial pattern of bias in the E-OBS dataset compared to higher quality data over the Alps, ELDAS domain and UK, expressed: the percentage of days that E-OBS data are more than 0.1 standard deviations below the higher quality data, *subtracted* from the percentage of days the E-OBS data are more than 0.1 standard deviation above the higher quality data. Thus, a positive value indicates that E-OBS data tend to be biased greater than the higher quality data, and vice versa. Precipitation is shown left, with UK top, Alps in the middle and ELDAS at the bottom.

Temperature (UK only) is shown right, with minimum temperature at the top, maximum temperature in the middle and mean temperature at the bottom.

Figure 6. Absolute error in different deciles for each comparison with existing datasets for precipitation (left) and temperature (right). In the left figure red is for the UK, green for the Alps and blue for ELDAS, in the right figure red is for minimum temperature, green for maximum temperature and blue for mean temperature. The box of absolute error shows the 0.25^{th}, median and 0.75^{th} percentile, the whiskers show the 0.05^{th} and 0.95^{th} percentile. Deciles are calculated for each grid separately.

Figure 7. Bivariate histograms showing the joint frequency distribution of cross validation error and interpolation standard deviation for precipitation (left) and minimum temperature (right). Both figures are on a log-log scale.

Figure 8. Spatial patterns of the percentage of interpolated data exceeding the lower (left) and upper (right) limits of the 95% confidence interval for precipitation (top) and minimum temperature (bottom) for all stations. Insets display histograms of the frequency of the over- or underestimation of the stations.