Interpolating observed global precipitation series

Juergen Grieser, j.grieser@gmx.de
The author is a director in the Model Development Department with Risk Management Solutions Ltd. (RMS)
This work, however, is neither done with RMS nor with the GPCC but in private initiative and with private means in the year 2005.

Database

The Global Precipitation Climatology Centre (GPCC) provided 3 station data sets to the author covering the period 1951 to 2000 with a maximum of 10% missing values:

1. 5,506 globally distributed time series of quality controlled and homogeneity tested observed monthly precipitation,
2. 3,210 additional observed German monthly precipitation time series and
3. 627 additional observed French monthly precipitation time series.

According to the GPCC these time series were tested for outliers.

The Food and Agriculture Organization of the UN (FAO) publicly provides

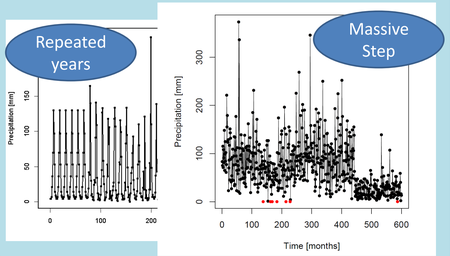
1. A set of 13,568 stations with observed monthly precipitation time series of various length (no QC) and
2. A set of 27,909 stations with observed long-term average monthly precipitations (no QC).

Fraction of common data from different sources within the GPCC database with equal (+/-1mm) observations in %.

	National	FAO	GHCN	CRU	Climat	Regional	GPCC	CPC
National	-	91.3	88.4	93.1	85.4	85.6	40.2	45.2
FAO	91.3	-	94.7	90.5	83.6	89.7	38.0	56.2
GHCN	88.4	94.7	-	98.8	84.3	74.2	43.1	56.9
CRU	93.1	90.5	98.8	-	87.3	87.7	39.0	53.6
Climat	85.4	83.6	84.3	87.3	-	74.6	42.2	56.3
Regional	85.6	89.7	74.2	87.7	74.6	-	41.6	39.8
GPCC	40.2	38.0	43.1	39.0	42.2	41.6	-	38.6
CPC	45.2	56.2	56.9	53.6	56.3	39.8	38.6	-

Visual Check

All time series are visually checked since it turned out that the homogeneity test of the GPCC was not applied to nearly 1,200 stations (since no reference series could be constructed). Obvious inhomogeneities are found in only 13 records.



Typical Inhomogeneities

Outlier Test

The GPCC claims to visually check the 2% lowest and highest observations (per station) for plausibility (Schneider et al. 2014). In case of the datasets provided to the author this would have been 224,232 visual checks.

The author estimated the probability that the highest observation per station and calendar month does not belong to the rest of the observations by

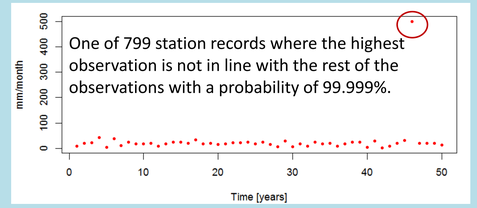
- Fitting a Weibull distribution to all but the highest observations,
- Perform a Mann-Kendall test of fit to ensure that the Weibull distribution reflects the observations,
- Calculate the probability that the highest observation does not belong to the rest, i.e.

$$p_x(x_x, N) \approx 1 - \exp\left(-N(1 - F(x_x))\right)$$

Outliers found:

# of outliers	Probability
7,728	99%
1,442	99.99%
799	99.999%

Those outliers need not to be unphysical. They can be extreme but true values that do not fit to the bulk of observations, i.e. due to atmospheric rivers, Vb weather type, etc.



Data Transformation

The following transformations are applied to the precipitation observations to test which variable performs best for interpolation.

Nr.	Name	transformation	inverse transformation
1	direct	y=x	x=y
2	abs. deviation	$y = x - \mu_x$	$x = y + \mu_y$
3	rel. deviation	$y = \frac{x-\mu_x}{\mu_x}$	$x = y\mu_y + \mu_y$
4	ratio	$y = \frac{x}{\mu_x}$	$x = y\mu_y$
5	standardization	$y = \frac{x-\mu_x}{\sigma_x}$	$x = y\sigma_y + \mu_y$
6	log1	$y = \ln(x + 0.5)$	$x = \exp(y) - 0.5$
7	log2	$y = \ln(x + 1)$	$x = \exp(y) - 1$
8	standardized log1	$y = \ln\left(\frac{(x+0.5)-\mu_{x+0.5}}{\sigma_x}\right)$	$x = (\exp(y)\sigma_y + \mu_{y+0.5}) - 0.5$
9	standardized log3	$y = \ln\left(\frac{(x+1)-\mu_{x+1}}{\sigma_x}\right)$	$x = (\exp(y)\sigma_y + \mu_{y+1}) - 1$

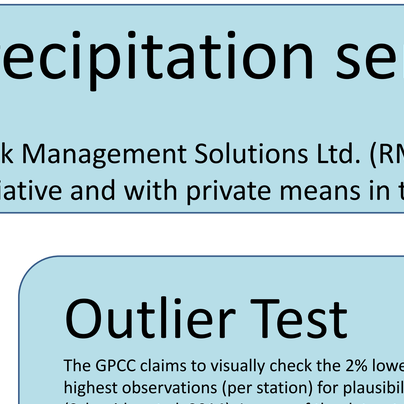
Interpolation Methods

1. Shepard's Method (original version in spherical coordinates)
2. Shepard's Method as installed at the GPCC by David Legates
3. Shepard's Method as altered by the GPCC (Becker et al. 2013)
4. Ordinary Kriging
5. Local Station Correlation (LSC).

All these methods use spatial scales. Shepard's length scale is solely a function of station density. The exponential variogram function used with Kriging uses a global scale as function of spatial variability as observed at station locations. LSC uses individual length scales for each station estimated by a station-specific correlation length.

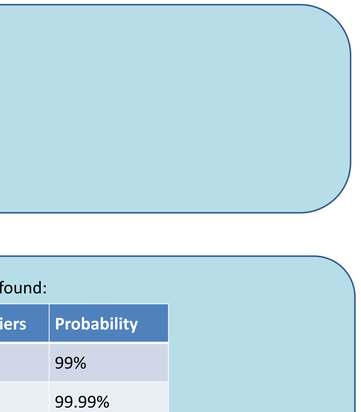
$$\rho^2(d) = \exp\frac{-d}{\lambda} \Leftrightarrow \lambda_i = \frac{\sum_{j=1}^{N} d_{ij}^2}{\sum_{j=1}^{N} d_{ij} \ln \rho_{ij}}$$

λ= correlation length, d= distance, ρ= correlation, j=1...N are the N neighbours of station i.



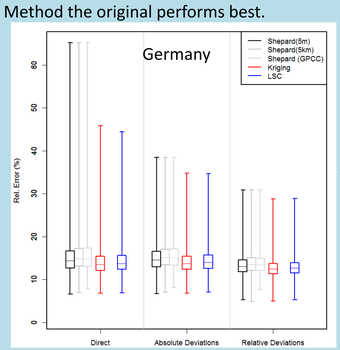
Station-specific correlation length [km].

Locally explained variance by closest station.

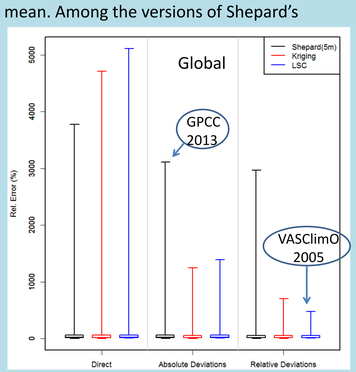


Comparison

Jack-knifing is used for the comparison of the interpolation methods and transformations. Station-specific bias, RMS and relative errors are calculated. Differences between interpolation methods are less important than between transformations. Shepards method performs worst for all transformations including the long-term mean. Among the versions of Shepard's Method the original performs best.



Relative interpolation error for three interpolation methods and three transformations.



Base Climatology

The interpolation of relative deviations from the long-term mean provides best results and is used to produce a 50-year gridded global precipitation dataset.
The mean precipitation at the grid points has to be estimated in order to perform the back transformation from relative deviations at the grid points.
This is done by Ordinary Kriging of the long-term averages of the time series used plus the 27,909 long-term averages provided by FAO.

The GPCC long-term averages were not made available to the author. The data provided by FAO are not quality controlled.

The VASClimO Dataset

- The VASClimO dataset is published by the GPCC (Beck et al. 2005; Rudolf and Schneider 2005) as a result of the VASClimO project.
- The dataset is neither produced by means nor on behalf of the GPCC or the VASClimO project.
- Contrary to (Beck et al. 2005) and (Rudolf and Schneider 2005) it is not interpolated by ordinary Kriging.
- Contrary to (Beck et al. 2005) and (Rudolf and Schneider 2005) it is not based on quality-controlled long term means of the GPCC but the uncontrolled FAO data.
- Since 2013 the GPCC interpolates absolute deviations from the mean "as a result of the VASClimO Project". The VASClimO dataset, however, is interpolated using relative deviations with maximum relative errors about factor 6 lower than with the GPCC strategy.
- From the 9,343 stations used for the VASClimO dataset, 3,345 stations are in Germany, 789 stations in France and 5,209 stations in the rest of the world, i.e. 44% of the stations cover 0.7% of the area.
- If only the original 5,506 stations had been used, station density in France would have been the third highest, Germany would rank 19 from 124 contributing countries (for comparison: the UN has 193 member states). This author opposes the statement of the German Weather Service (DWD 2010) that it was necessary to use the German and the French data sets to fill a gap that would result otherwise.
- The outlier test showed that the GPCC is not applying the test strategy they pretend to apply (Hechler 1999, Schneider et al. 2014).