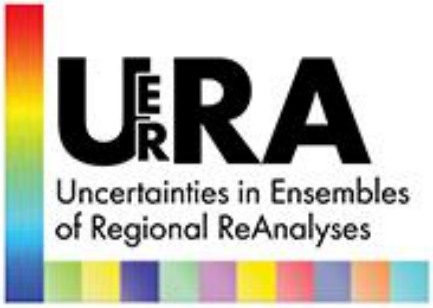
Seventh Framework Programme
Theme 6



Project: 607193 UERRA

Full project title:
Uncertainties in Ensembles of Regional Re-Analyses

Deliverable D1.14
Development of more comprehensive uncertainty estimations

WP no.:	1
WP leader:	URV
Lead beneficiary for deliverable:	UEA
Name of author/contributors:	Richard Cornes (KNMI) and Phil Jones (UEA)
Nature:	Other
Dissemination level:	PU
Deliverable month:	36
Submission date:	11 April 2017
Version nr:	1

1. Introduction

Emphasis has been placed in this project on the generation of more comprehensive estimates of uncertainty in the E-OBS dataset through the development of stochastic measures of uncertainty. These estimates of uncertainty are derived from multiple realizations of the daily gridded fields. Few gridded datasets provide estimates of uncertainty, and even fewer estimate this uncertainty from an ensemble of realizations. Notable exceptions at the regional scale are the work of Clark et al. (2006) and Newman et al. (2015) and at the global scale the HadCRUT4 dataset (Morice et al. 2012). The covariance of the gridded field in such datasets not only provides a more rigorous estimation of uncertainty in the gridded fields themselves but is particularly useful for generating uncertainty estimates in derived applications of the data. It should be stressed, however, that while the use of the term "ensemble" is used in these studies – as well as for the technique described here for the E-OBS dataset – this is quite a different estimation of uncertainty than is provided by an ensemble of climate model simulations (for example in the reanalysis data developed by UERRA), which use the same terminology. In the case of the model simulations the uncertainty is generated from changing the initial conditions of the model, whereas for the station-based interpolations described in this paper the uncertainty is a stochastic measure relative to the method of interpolation, and is ultimately determined by the density of input station data, and the suitability of the covariates used to effect the interpolation.

2. The New E-OBS Daily Gridding Methodology

In Deliverable D1.10 the regression-kriging method was introduced as a feasible way of interpolating the monthly values of temperature and precipitation in E-OBS. However, it became clear in the application of this method for the daily data that compound uncertainty between the daily and monthly fields, and between the regressed “spatial trend” and the kriged regression residuals, led to ambiguity in the estimates of uncertainty in the final daily grids. We have therefore opted for a single Generalized Additive Model (GAM) for each day. This gridding technique when applied to the E-OBS data is now considered stable, and a full 1950-2015 daily interpolation completes in a matter of hours, when run on the high-performance cluster at ECMWF. Efforts are currently under way to make this dataset operational.

2.1 The statistical model

In the general case, the GAMs used in this new interpolation of E-OBS follow those described by Wood et al. (2015, 2016), and evaluate the relationship between the temperature or rainfall response variable y_i to one or more predictors x_{ji} as

$$y_i = \sum_{j=1}^{n} f_j(x_{ji}) + \epsilon_i, \quad y_i \sim N(0, \sigma^2)$$

where for n smoothing parameters, f_j are estimated for each of j parametric vector covariates as part of the model fitting. The optimal fitting of the smooth functions f_j is achieved using Restricted Marginal Likelihood (REML), which tends to suffer less from under-smoothing than the traditional GCV optimisation (Hastie & Tibshirani, 1990).

The interpolation proceeds as two stages: long-term climatological averages for each month calculated over the period 1961-90 are initially modelled using environmental parameters, and daily

values are then modelled through a combination of latitude and longitude coordinates plus the background climatological field interpolated to the daily station locations. Specifically, the climatological variables are evaluated for each month as

$$y_i = f_1(lon_i, lat_i) + f_2(alt_i)$$

where f_1 takes a reduced-rank thin-plate spline basis using latitude and longitude parameters, and $f_2$is a cubic spline response of altitude. As in the earlier versions of E-OBS, station-based altitude values are used for model-fitting, while GTOPO30 Digital Elevation Model (DEM) values of altitude are used for the interpolation. Although GTOPO30 has been superseded by GMTED2010 (Danielson & Gesch, 2011), we continue to use GTOPO30 in this analysis to ensure that the comparison of the new version of E-OBS against the earlier version is restricted to the interpolation method, without the introduction of a confounding effect of different topographic data. We have also tested the potential for improved model fitting through the incorporation of additional co-variates to the climatological models, which take cubic spline smoothing bases. In the case of the temperature variables we have included distance from the nearest coast and the topographic position index (TPI). Coastal proximity was calculated using a relatively coarse 1:110m resolution coastline so as to provide a measure to the nearest large body of water rather than minor coastal features (Daly et al. 2008). TPI is calculated as the difference between a cell and the mean of the nearest eight cells. It is intended to provide a measure of small scale features such as frost hollows. In the case of rainfall we have also included measures of slope (α) and aspect (Θ) measured through the horizontal components p and$q$, where

$$p = -\cos\alpha \sin\Theta,\ q = -\sin\alpha \sin\Theta$$

after Hutchinson (1998). These topographic parameters were calculated using the ~1km GTOPO30 data.

Since the models used in this interpolation are additive, the individual model terms can easily be extracted. These terms all represent physically plausible mechanisms of the temperature and precipitation climatology of Europe. The longitude/latitude model term represents the expected zonal pattern for temperature and meridional pattern for rainfall (Figure 1). During the winter months and to some extent the autumn, the latitudinal pattern of the temperature smoothing parameters is distorted along the Atlantic seaboard of Europe through sea-surface temperature effects. The variation in maximum and minimum daily temperature (not shown) broadly follow the results for mean temperature. In the case of the temperature variables, the altitude term (Figure 2) shows the expected lapse rate, which is greater than or less than the Dry Adiabatic Lapse Rate (DALR) depending on the month. In the case of rainfall this model-term indicates an increase in rainfall with altitude, which occurs a rate of ~2mm km^1 up to ~1500m, but drops significantly beyond that altitude after which an increase is observed. This drop at ca. 2000m is possibly related to rain-shadow effects at higher altitudes, although it should be noted that there are few stations at that relatively high altitude. Coastal proximity has a strong warming effect on the temperature climatology during the autumn and winter months, but only up to ca. 100km distance from the coast, which is the same distance over which the effect was considered to be significant in the interpolation of temperature across the UK by Jarvis & Stuart (2001); during the spring and particularly the summer months the expected cooling effect of coastal proximity is captured by this parameter. The zonal and meridional slope and aspect vectors (Qvec and Pvec respectively) indicate a complex relationship between topography and the rainfall climatology. Strongly west-facing

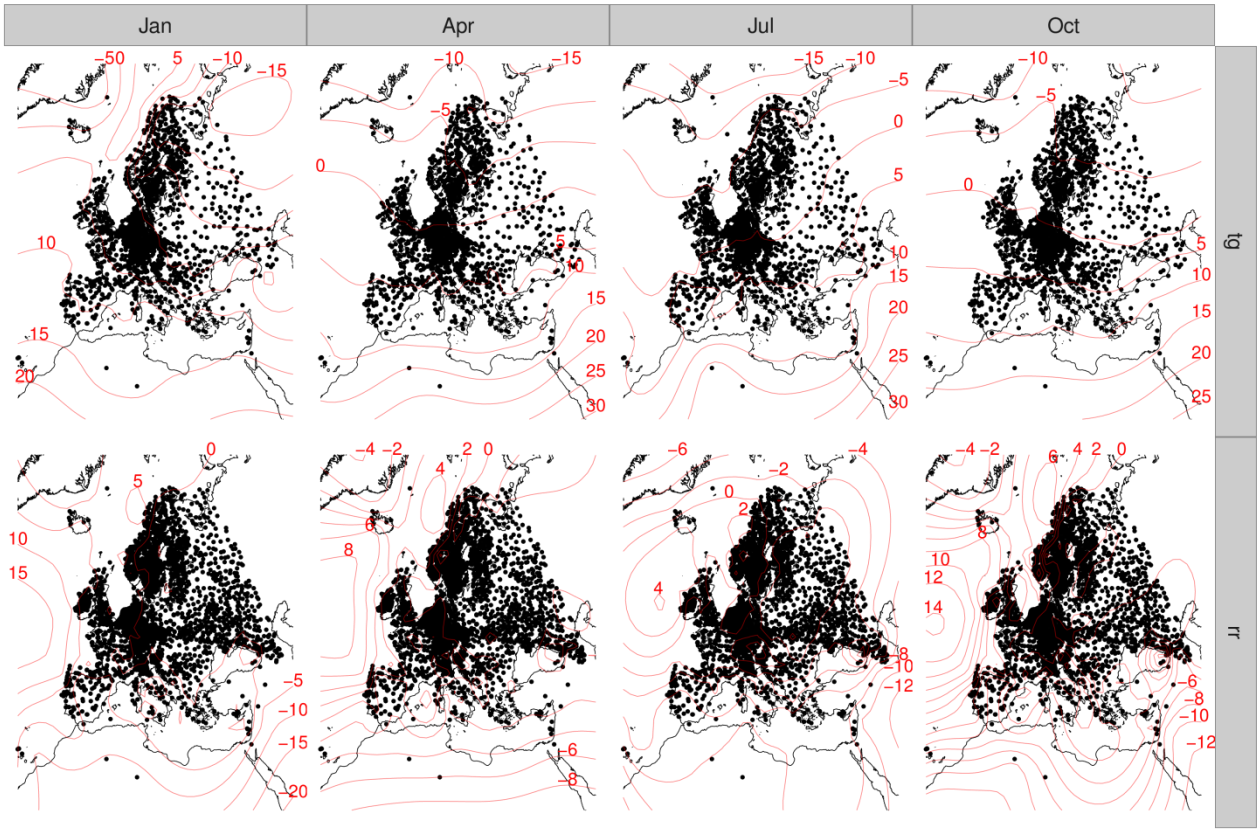


Figure 1. The longitude/latitude smoothing parameters from the climatological GAMs for four months of the year, and for the variables of daily mean temperature (tg) and daily total rainfall (rr). The units are in °C and √mm for temperature and rainfall respectively.

slopes are associated with higher rainfall totals, and weakly so for slopes with a strong east-facing gradient. As would be expected the meridional vector has less of an effect on the rainfall climatology, although there is a notable reduction in rainfall where Pvec=0.2, although there are few stations exhibiting such a strong southerly inclination. It should be noted, however, that both of these vectors (Pvec and Qvec) have a weaker influence on rainfall than altitude.

The daily interpolation of temperature and rainfall is achieved using a GAM of the form

$$y_i = f_3(lon_i, lat_i) + f_4(clim_i)$$

where $clim_i$ are the local climatological values estimated from the climatological models for the respective month of the year. In the case of daily rainfall totals, the model is fitted to square root transformed values of y_i to remove some of the skewness of the data. Tweedie distributions were tested since these models are suitable where the variance is proportional to the mean and have shown to be optimal for modelling rainfall values (Hasan & Dunn 2011). However, models using these distributions often failed to converge, particularly for days with a high number of zero values – a problem of greater frequency during the summer months.

In these models the model residual ϵ_i is assumed to be normally distributed and importantly without spatial autocorrelation. The longitude/latitude smoothing parameters are used in these models to account for this spatial autocorrelation. As discussed by Hefley et al. (2016) this captures

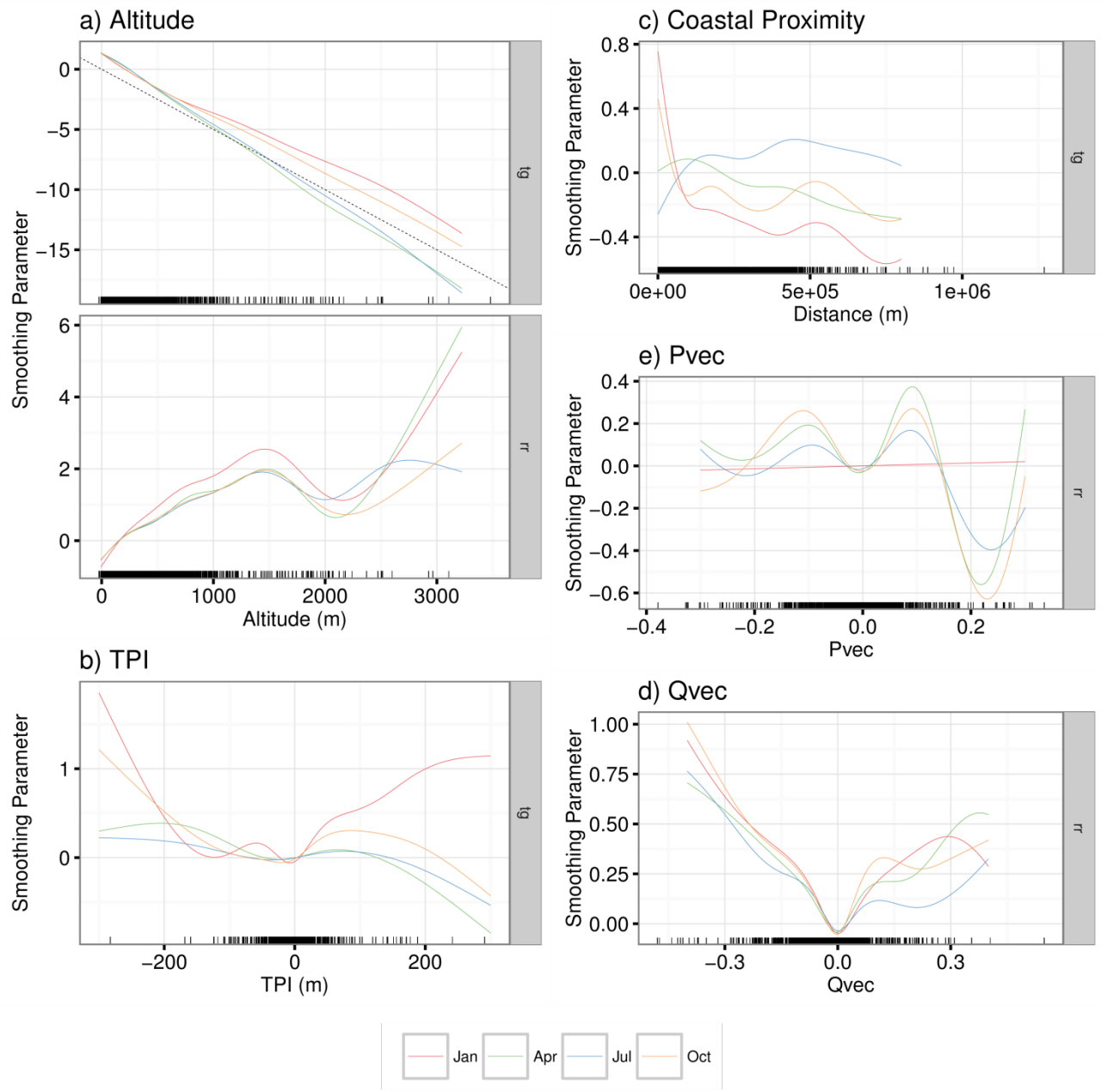


Figure 2. The smoothing parameters with cubic spline bases from the mean temperature (tg) and rainfall (mm) climatological GAMs for four months of the year. The black bars represent the station values that have been used in the models.

first-order spatial autocorrelation of the variable and hence the function f_3 describes autocorrelation in the mean of the distribution. Second-order functions captures autocorrelation in the covariance of the distribution and include the geoadditive models of Kammann & Wand (2003), whereby spatial autocorrelation is modelled using a pre-defined covariance function. In both the daily and climatological models, longitude and latitude were transformed from a regular grid (WGS80) to a Lambert Equal Area projection (EPSG:3035). This was favoured over the rotated-pole transformation of interpolation coordinates used in the previous versions of E-OBS, which suffers from distortion when converted back to regular coordinates.

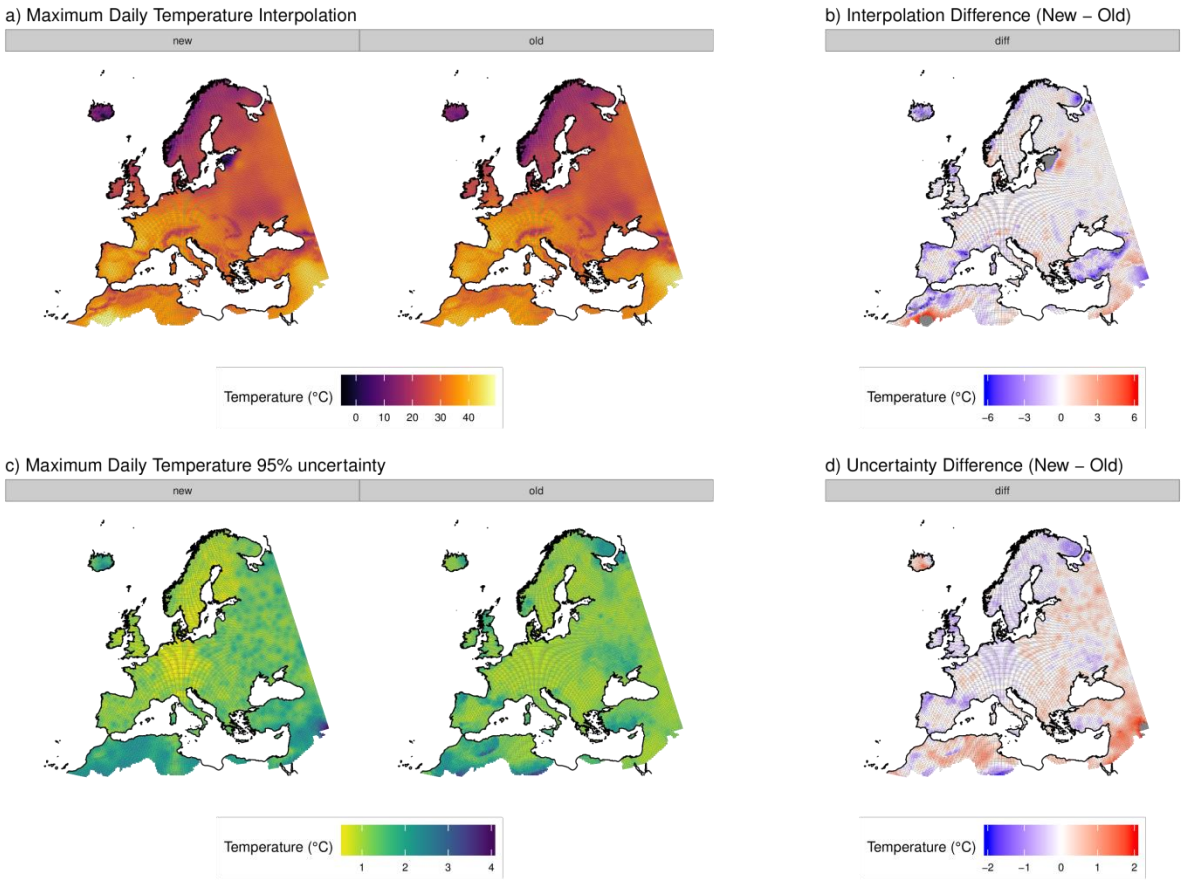


Figure 3. The maximum daily temperature interpolations for 4th August 2003 (a) between the current operational version of E-OBS (version 14.0, old) and the GAM-derived interpolation (new). Also included are the differences between the interpolations (new minus old, b), and 95%ile uncertainty values (c) and the difference between the uncertainty values. The two interpolation versions have been masked to a common grid. The "new" interpolation in a) is the mean across the 100 grid realizations.

Although this interpolation procedure consists of two stages – in a similar vein to the previous E-OBS method - it marks a significant departure from the former approach. Rather than using the interpolated monthly means to constrain the daily values, topographic effects are incorporated by using the interpolated climatological values as a covariate in the GAM at the daily resolution. This follows the approach of Masson & Frei (2014) and Hiebl & Frei (2016, 2017), who found that the incorporation of topographic effects to daily rainfall data across the Alps and Austria via a function of the interpolated climatological variable to be an improvement over the more usual interpolation of climatological values and daily proportions separately. It should be noted, however, that these studies used this approach under an external-drift kriging procedure, and hence the climatological values were incorporated as a linear response. Here non-linearity is permitted.

2.2 The uncertainty estimates

In the initial construction of the E-OBS dataset the possibility of generating an ensemble of grids for each day was explored, with the spread across the ensemble providing a measure of uncertainty in the interpolated values (Haylock et al., 2008). Due to the significant computational burden that this procedure entails – at least for the methods used in that interpolation – that approach was rejected in favour of a single uncertainty value for each day, which was derived from a combination of the monthly-climatological uncertainty estimates and daily kriging uncertainty, derived using the technique developed by Yamamoto (2000).

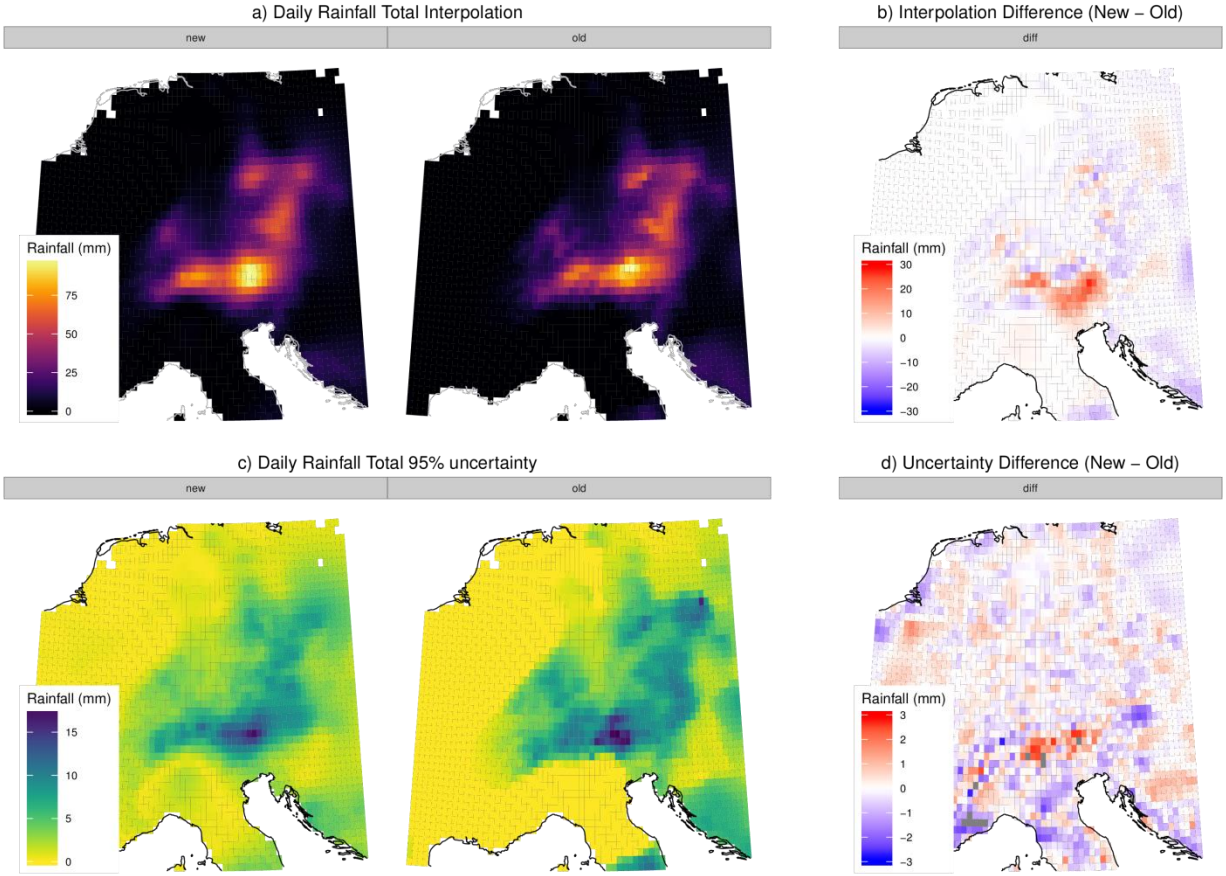


Figure 4. As Figure 3, but for the heavy rainfall event of 1st June 2013 across central Europe.

In this new version of E-OBS we estimate uncertainty in the daily interpolation by drawing simulations from the posterior distribution of the fitted model coefficients β, following the method described by Wood (2006) and Marra & Wood (2012). The Bayesian reasoning used in this approach is essentially the same approach used in the thin-plate splines of the earlier version of E-OBS, but is extended for use in the penalized GAMs used in this interpolation. This is achieved by simulating 100 sets of random vectors from β. Under this approach the simulations are conditional on the estimated smoothing parameters $\hat{\lambda}$ (Wood, 2006). From the distribution of these simulations the confidence intervals can be calculated by taking quantiles. However, as described by Wood (2006), while simulations are provided that are *conditional* on $\hat{\lambda}$ using this method, simulations that are *unconditional* on $\hat{\lambda}$ are desirable. The techniques suggested by Wood (2006) for overcoming this failed for certain daily models and therefore to ensure consistency across the time period, we continue to use simulations that are *conditional* on $\hat{\lambda}$. In general this has only small effects on the confidence interval of the fitted splines but can lead to a smaller confidence interval in the prediction than expected for a given quantile(Wood, 2006).

1.3 Demonstration of the E-OBS gridded fields for two extreme events

In this section we demonstrate the daily rainfall and temperature interpolation, along with the new uncertainty estimates, for two extreme events: the heavy rainfall event of 1st June 2013 (see http://cib.knmi.nl/mediawiki/index.php/Central_European_flooding_2013) and the climax of the 2003 European heatwave, on the 4th August 2003.

The interpolation for the 2003 heatwave (Figure 3) is broadly similar between the two E-OBS versions, although the new interpolation is locally warmer across many central European regions and significantly cooler across most mountainous regions of Europe. The uncertainty estimates in the new interpolation depict much more closely the location of stations used in the interpolation, where uncertainty in much reduced.

The interpolation of the heavy rainfall event of 1st June 2013 (Figure 4) tends to be smoother than the earlier version of E-OBS, with less gridbox-to-gridbox variation. However, the highest rainfall total is higher in the new interpolation. The increased spatial variance in the old interpolation results from the localized kriging technique used in previous versions of E-OBS, which restricts the number of stations to between 4 and 20, and the averaging of nine higher resolution grid-points to constitute the final ~25km grid cells. However, this higher variance gives a false impression of the variability of the daily rainfall field since the interpolation is not able to resolve rainfall patterns, at best, below ~25km. This spurious variance is evident in tests conducted as part of UERRA in the comparison of E-OBS against the high-resolution gridded datasets constructed by most National Meteorological Services (NMS) across Europe. Mean-error statistics, and percent-correct statistics are at least as comparable in the new version of E-OBS compared to earlier versions, and in certain cases the new interpolation is closer to the NMS data.

Acknowledgments

We are grateful to Christoph Frei for his advice in the development of this new version of E-OBS.

References

Clark, M. P., Slater, A. G., Clark, M. P., & Slater, A. G. (2006). Probabilistic Quantitative Precipitation Estimation in Complex Terrain. *Journal of Hydrometeorology*, 7(1), 3–22. http://doi.org/10.1175/JHM474.1

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H. & Pasteris, P. P. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, 28(15), 2031–2064. http://doi.org/10.1002/joc.1688

Danielson, J. J., & Gesch, D. B. (2011). Global multi-resolution terrain elevation data 2010 (GMTED2010).

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. Chapman & Hall/CRC.

Hasan, M. M., & Dunn, P. K. (2011). Two Tweedie distributions that are near-optimal for modelling monthly rainfall in Australia. *International Journal of Climatology*, 31(9), 1389–1397. http://doi.org/10.1002/joc.2162

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research*, 113(D20), D20119. http://doi.org/10.1029/2008JD010201

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., & Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3), 632–646. http://doi.org/10.1002/ecy.1674

Hiebl, J., & Frei, C. (2016). Daily temperature grids for Austria since 1961—concept, creation and applicability. *Theoretical and Applied Climatology*, 124(1–2), 161–178. http://doi.org/10.1007/s00704-015-1411-4

Hiebl, J., & Frei, C. (2017). Daily precipitation grids for Austria since 1961—development and evaluation of a spatial dataset for hydroclimatic monitoring and modelling. *Theoretical and Applied Climatology*, 1–19. http://doi.org/10.1007/s00704-017-2093-x

Hutchinson, M. F. (1998). Interpolation of Rainfall Data with Thin Plate Smoothing Splines - Part I: Two Dimensional Smoothing of Data with Short Range Correlation. *Journal of Geographic Information and Decision Analysis*, 2, 168–185.

Jarvis, C. H., Stuart, N. (2001). A Comparison among Strategies for Interpolating Maximum and Minimum Daily Air Temperatures. Part I: The Selection of "Guiding" Topographic and Land Cover Variables. *Journal of Applied Meteorology*, 40(6), 1060–1074. http://doi.org/10.1175/1520-0450(2001)040<1060:ACASFI>2.0.CO;2

Kammann, E. E., & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1), 1–18. http://doi.org/10.1111/1467-9876.00385

Marra, G., & Wood, S. N. (2012). Coverage Properties of Confidence Intervals for Generalized Additive Model Components. *Scandinavian Journal of Statistics*, 39(1), 53–74. http://doi.org/10.1111/j.1467-9469.2011.00760.x

Masson, D., & Frei, C. (2014). Spatial analysis of precipitation in a high-mountain region: exploring methods with multi-scale topographic predictors and circulation types. *Hydrology and Earth System Sciences*, 18(11), 4543–4563. http://doi.org/10.5194/hess-18-4543-2014

Morice, C. P., Kennedy, J. J., Rayner, N. A., & Jones, P. D. (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research Atmospheres*, 117(8), D08101. http://doi.org/10.1029/2011jd017187

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., and Arnold, J. R. (2015). Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. *Journal of Hydrometeorology*, 16(6), 2481–2500. http://doi.org/10.1175/JHM-D-15-0026.1

Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1), 139–155. http://doi.org/10.1111/rssc.12068

Wood, S. N., Li, Z., Shaddick, G., & Augustin, N. H. (2016). Generalized additive models for gigadata: modelling the UK black smoke network daily data. *Journal of the American Statistical Association*, 1–40. http://doi.org/10.1080/01621459.2016.1195744

Yamamoto, J. K. (2000). An alternative measure of the reliability of ordinary kriging estimates. *Mathematical Geology*, 32(4), 489–509. http://doi.org/10.1023/A:1007577916868