Simulating space-time uncertainty in continental-scale gridded precipitation fields for agrometeorological modelling

Allard de Wit and Sytze de Bruin
Centre for geo-information, Wageningen UR
P.O. Box 47, 6700 AA, Wageningen
Tel.: + 0031 317 474761; Fax: + 0031 317 419000
Allard.dewit@wur.nl; Sytze.debruin@wur.nl

Abstract
Previous analyses of the effects of uncertainty in precipitation fields on the output of EU Crop Growth Monitoring System (CGMS) demonstrated that the influence on simulated crop yield was limited at national scale, but considerable at local and regional scales. We aim to propagate uncertainty due to precipitation in the crop model by Monte Carlo sampling of the precipitation field. We use an error model fitted to a highly accurate precipitation dataset (ELDAS) which was available for the year 2000. Our error model consisted of two components. The first is an additive component generating precipitation residues over the entire spatial domain. The residues are generated by quantile-based back transformation of standard Gaussian fields using a set of histograms for different CGMS precipitation bins. The second component is multiplicative and generates binary rain/no-rain events on locations where the CGMS precipitation records report nil precipitation. Our results demonstrate that the model generates realistic patterns of precipitation and reproduces the histograms of the reference precipitation dataset well. A remaining problem is the inability to model prolonged dry spells which is due to our model choice. The precipitation realizations were used as input in a crop growth model. The first results indicate that the uncertainty in precipitation is sufficient to sustain divergence in the soil moisture ensemble, but not in the leaf area index ensemble.

Keywords: precipitation, error model, multiple realisations, Gaussian field, crop model

1 Introduction
Analyses of the effects of uncertainty in the precipitation data for the year 2000 on the output of a spatially distributed crop model called Crop Growth Monitoring System (CGMS) demonstrated that the influence on simulated crop yield was limited at national scale, but considerable at local and regional scales (Wit et al. 2005). The current paper aims to contribute to the development and parameterisation of a method for generating realistic ensembles of precipitation residuals representing the uncertainty in the CGMS precipitation dataset or any other routinely created gridded precipitation dataset. Such methods are needed for assessing error propagation through complex non-linear models such as crop growth models or hydrological models (Beven and Freer 2001) and can be used in a data assimilation approach using, for example, an ensemble Kalman filter.

Most methods for representing uncertainty in precipitation inputs consist of generating an ensemble of precipitation inputs through some stochastic process. The ensemble generation approaches are employed in different situations using often similar (geo)statistical methods but with different emphasis. 1) Weather generator applications which generate time-series of precipitation on single locations or on multiple locations with or without spatial dependence...
(Wilks and Wilby 1999: Yang et al. 2005). 2) Downscaling of coarse numerical weather prediction (NWP) precipitation fields to point locations or to higher spatial and/or temporal resolution fields with realistic properties, often applied within a precipitation nowcasting framework (Bates et al. 1998: Charles et al. 2004: Clark et al. 2004: Seo et al. 2000). 3) Interpolating from weather stations to ungauged locations or for simulating uncertainty in a gridded precipitation product (Carpenter et al 2004: Lanza 2000: Pardo-Igúzquiza et al. 2006).

Generating ensembles of spatially distributed precipitation estimates requires that either the spatial structure of the precipitation fields itself has to be taken into account (Lanza 2000: Seo et al. 2000), or the spatial structure of a residual error field (Kyriakidis et al. 2004). One of the difficulties in implementing these schemes is that a model (often a variogram) is needed describing the spatial structure. Seo et al. (2000) modelled the rainfall structure directly and these models were selected based on the type of storm and the expected correlation structure. Apart from the operational problems of selecting a spatial model event by event, the authors also noted that the selected variograms are sometimes not very representative of the generally nonhomogeneous structures of observed precipitation. Similarly, the approach of Lanza (2000) needs an estimate of the spatial covariance structure and the fraction of rain/no-rain areas to describe the structure of the rainfall fields. Kyriakydis et al. (2004) did not need to specify a model for spatial precipitation structure directly but rather a model for the residual error fields.

In this paper, we propose a method for generating residual error fields for daily precipitation data over large areas using a relatively simple methodology. The method takes into account spatial correlations in the ensemble. It should reproduce the input statistics (mean and variance) and presents a pragmatic solution that can be parameterized relatively easily. Moreover, we present some first results obtained by running an ensemble of crop models using an ensemble of rainfall inputs.

2 Data

2.1 CGMS meteorological database

The CGMS meteorological database contains weather information starting in the 1970's and is continuously updated with weather information. This long time-series of weather data is important for retrospective analyses of crop stress situations and validation of crop yield forecasts. The information in the database is currently derived from about 2500 weather stations over Europe, Turkey and the Maghreb. The total number of stations varies over time as a result of stations being discontinued or new ones established.

An interpolation routine is applied to estimate weather variables for each 50x50 km grid cell (Voet et al. 1994). Each cell receives values for temperature, radiation, vapour pressure, evapotranspiration and wind speed using inverse distance weighting of the so-called “meteorological distance”. This meteorological distance is not only based on the Euclidean distance between the cell centre and the weather station, but also accounts for factors like altitude, distance to coast and the existence of climate barriers (e.g. mountain ridges, water bodies) between the grid cell and the weather stations. In case of rainfall a grid cell receives the value of the weather station with the smallest meteorological distance from the grid cell because averaging of rainfall values was found to decrease the interpolation accuracy.
2.2 ELDAS precipitation data

The ELDAS precipitation database consists of daily precipitation values on a 0.2° grid over Europe for the period 1 October 1999 until 31 December 2000 (Rubel; Hantel 2001: Rubel et al. 2004). The precipitation values were interpolated using block kriging based on more than 20,000 bias-corrected rain gauge measurements. The collection of these rain gauge measurements was a one-off activity and no update is to be expected in the near future. Validation has demonstrated that systematic measurement errors for over 90% of the number of stations are within 1 mm/day. Given the sheer volume of rain gauge measurements that were used to generate this database, it will give a far better estimate of the true rainfall patterns compared to the estimates in the CGMS meteorological database. We use the ELDAS database as a reference for modelling the error structure in the CGMS precipitation fields. The ELDAS precipitation database was converted to the 50x50 km CGMS grid by taking the average precipitation of ELDAS cells within a CGMS grid cell.

3 Method

3.1 Conceptual modelling

We consider \( \text{Prec}_{\text{ELDAS}}(x,t) \) to be the spatial average precipitation at grid cell \( x \) and day \( t \), \( \text{Prec}_{\text{CGMS}}(x,t) \) the precipitation as recorded in the CGMS system, and \( \varepsilon(\cdot) \) a random spatially and/or temporally correlated residual. Our spatio-temporal error model for the data concerned is then given by:

\[
\varepsilon_{o}(x,t) = \text{Prec}_{\text{ELDAS}}(x,t) - \text{Prec}_{\text{CGMS}}(x,t)
\]

(1)

Note that no statement is made about \( \varepsilon(\cdot) \) having zero mean, as \( \text{Prec}_{\text{CGMS}}(\cdot) \) might be biased (see section 2.3.2).

The basic assumption underlying our method is that the residuals \( \varepsilon_{o}(x,t) \) can be transformed to standard (marginal) normality by some transform function \( f(\cdot) \) and that correlated standard Gaussian fields can be back transformed to residual precipitation fields by the inverse of that function, i.e. \( f^{-1}(\cdot) \). We thus obtain the following model for the random residual precipitation field, \( \varepsilon_{M}(\cdot) \):

\[
\varepsilon_{M}(x,t) = f^{-1}(\delta(x,t), \text{args})
\]

(2)

Where \( \delta(\cdot) \) denotes a correlated standard Gaussian random field and \( \text{args} \) are any required additional arguments. Section 3.2 (below) explains that the zero residuals in case of zero CGMS precipitation require additional modelling.

Once the transformation functions \( f(\cdot) \) and \( f^{-1}(\cdot) \) and the random function generating \( \delta(\cdot) \) are configured using the observed \( \varepsilon_{o}(x,t) \), we are able to generate multiple realizations of \( \varepsilon_{M}(\cdot) \) using any standard Gaussian simulation algorithm. By summing the simulated residuals to observed CGMS precipitation data the required ensemble traces are obtained.
3.2 Normal score transformation

We used quantile based normal score transforms for a series of $\text{Prec}_{\text{CGMS}}(\cdot)$ intervals. The dataset for the year 2000 was divided into 13 CGMS precipitation intervals and for each of these a histogram of the observed residuals $\epsilon_O(\cdot)$ was produced. Next, the normal scores of the observed residuals and 13 transformation tables were obtained by finding the $z$ scores of a standard Gaussian distribution corresponding to quantiles of the observed cumulative distributions. The computations were done using the GSLIB program nscore (Deutsch; Journel 1998).

The first bin $\text{Prec}_{\text{CGMS}}(x,t) = 0$ required additional processing, because the transformation algorithm could not properly transform the large number of zeros (i.e. $\epsilon_O(x,t) < 1$ mm) in the residuals. Therefore, we introduced a multiplicative —spatially correlated— indicator variable $i(x,t)$ to treat the $\epsilon_O(x,t) < 1$ mm data given $\text{Prec}_{\text{CGMS}}(x,t) = 0$ mm separately. Hence, in the first bin only the residuals $\epsilon_O(x,t) \geq 1$ mm are handled by a normal score transform and Eq. [2] is modified into:

$$
\epsilon_M(x,t) = \begin{cases} 
  i(x,t) \cdot f^{-1}(\delta(x,t), \text{Prec}_{\text{CGMS}}(x,t)) & \text{if } \text{Prec}_{\text{CGMS}}(x,t) = 0 \\
  f^{-1}(\delta(x,t), \text{Prec}_{\text{CGMS}}(x,t)) & \text{if } \text{Prec}_{\text{CGMS}}(x,t) > 0
\end{cases} 
$$

with

$$
i(x,t) = \begin{cases} 
  \text{null} & \text{if } \text{Prec}_{\text{CGMS}}(x,t) > 0 \text{ mm} \\
  1 & \text{else, if } \epsilon_O(x,t) \geq 1 \text{ mm} \\
  0 & \text{otherwise}
\end{cases}
$$

In the current work, dependence of $i(x,t)$ on $\text{Prec}_{\text{CGMS}}(\cdot)$ $|\text{Prec}_{\text{CGMS}}(x,t) > 0$ and $\delta(\cdot)$ is not considered.

3.3 Variogram modelling

3.3.1 Normal score transformed data

Normal score transformed residuals $\epsilon_O(\cdot)$ were computed for all $\text{Prec}_{\text{CGMS}}(x,t) > 0$ mm and for $\text{Prec}_{\text{CGMS}}(x,t) = 0$ mm with $\epsilon_O(x,t) \geq 1$ mm. The thus obtained data was exhaustively sampled in the spatial domain and thrice-monthly, i.e. on the 1st, 11th and 21st of a month in the temporal domain to determine the experimental variogram of the spatially autocorrelated Gaussian variable $\delta(\cdot)$. A variogram model was interactively fitted through the data points (average of all dates). The fitted function was forced to attain the theoretical unit sill corresponding to a standard Gaussian field.

3.4 Checking two-point normality

The normal score transformed residuals $\epsilon_O(\cdot)$ are by construction univariate normally distributed, but the nscore transform does not impose multivariate normality on $\epsilon_O(\cdot)$.
Nonetheless, the random function $\delta(\cdot)$ (Eq. 4), which is configured on $\varepsilon_0'(x,t)$, assumes the two-point distribution of any pair of values at different locations to be Gaussian. To check the consequences of this assumption for the intended use of the data, we employed a procedure given in Goovaerts (1997, pp. 271-275) and Deutsch and Journel (1998, pp. 142-144) which consists in graphically comparing experimental and Gaussian model-induced indicator variograms of the normal score data at different $p$-quantiles of the cumulative distribution.

The graphical comparisons of the experimental and Gaussian model-induced indicator variograms of the normal score data demonstrated that the normal score data are approximately bivariate normally distributed. Thus we decided to accept the assumption of two-point normality in space for the purpose of our study.

![Figure 1 Flow diagram of the ensemble simulation approach.](image)

### 3.5 Simulation of residual fields

Figure 1 shows how we implemented our simulation method. If $\text{Prec}_{CGMS}(x,t) > 0$ mm, then the right-hand branch of the flow diagram suffices, i.e. unconditional standard Gaussian simulation followed by a back transform. In the other case, unconditional indicator simulation is used to model the event of a positive residual $\varepsilon_t(x,t)$ given $\text{Prec}_{CGMS}(x,t) = 0$ mm. Note, however, that both branches are always executed and that the conditions $\text{Prec}_{CGMS}(x,t) > 0$ mm or $\text{Prec}_{CGMS}(x,t) = 0$ mm are handled by post processing (lower box in Figure 1).
The Gaussian and indicator simulations were performed using the public domain sequential simulation programs \textit{sgsim} and \textit{sisim} included in GSLIB 2 (Deutsch and Journel, 1998). The back transforms and post processing were performed by the LINT2 module in the TTUTIL Library (Kraalingen and Rappoldt 2000) and the Python scripting language. A set of 100 alternative realisations of daily precipitation was generated by adding back transformed simulated residuals to the CGMS precipitation data.

\subsection*{3.5.1 Evaluation of precipitation realizations}
We evaluated the realisations of the precipitation fields on four different aspects. Firstly, the reproduction of the histograms of the ELDAS precipitation was checked by quantile-quantile (Q-Q) plots of ELDAS precipitation against five precipitation realisations for two selected days. Secondly, variogram reproduction of the ELDAS precipitation fields was evaluated using the same days and realisations that were used for histogram reproduction. Thirdly, the rainfall temporal intermittency characteristics for both dry and wet periods were compared for 6 representative sites for the CGMS precipitation, ELDAS precipitation and 25 realisations.

\section*{4 Results}

\subsection*{4.1 Precipitation realizations}
The Q-Q plots of Figure 2 show that for 21 November 2000 the distributions were nearly identical up to 30 mm, but the 5 precipitation realisations overrepresented the number of precipitation occurrences greater than 30 mm. Similarly, for 11 July 2000 the distributions were nearly identical up to 15 mm but the realisations contained too many precipitation occurrences greater than 15 mm.

![Figure 2 Quantile-Quantile plots of precipitation quantiles of five realisations versus the observed ELDAS precipitation quantiles.](#)

The variograms of the same 5 realisations at 21 November 2000 and 11 July 2000 (Figure 3) demonstrate that the overall variance (the sill) of the realisations can be both smaller and larger than the ELDAS variance (grey line), but that the range of the spatial correlation was similar to...
the ELDAS reference dataset. However, at small ranges all realisations had larger variability in precipitation values than ELDAS.

The intermittency characteristics of the dry periods were determined for six representative sites located in areas with major agricultural production (Only two sites are shown: Figure 4). The results for South-Spain (Andalusia) demonstrate that the proposed simulation approach was not able to reproduce the dry period statistics well. The characteristic long dry summer in Andalusia with dry periods as long as 130 days are not reproduced in the realisations. This is a result of the fact that the simulation approach can generate a precipitation residual for zero CGMS precipitation, depending on the outcome of the independent indicator simulation. This property of the simulation approach “breaks” apart the very long dry periods during the Andalusian summer. Similar effects can be observed in the results for Northern Spain and Southern France albeit less pronounced.

The results for the sites located in more temperate climate regions (Northern-France, Central-Germany, Denmark) demonstrated that the simulation approach performed well in reproducing the dry period statistics. Still, the dry period lengths in the realisations are, on average, too short.

The intermittency characteristics of the wet periods were determined for the same six sites (only two are shown in Figure 5). It can be observed that, compared to the dry period length, the simulation approach performs much better in reproducing the wet period intermittency characteristics. Nevertheless an increase in the number of events greater or equal to one day can be observed for nearly all sites.
4.2 Crop model simulations

Figure 6 shows some first results from running the CGMS crop model using an ensemble of precipitation inputs. The initial soil moisture states were created by sampling from a random variable with mean half way between wilting point and field capacity and the standard deviation equal to 20% of the soil moisture range between wilting point and field capacity.

The results for south Spain demonstrate that the average soil moisture level increases slightly up to day 80, the a strong decline in soil moisture levels occurs because the crop starts to draw water from the soil and no precipitation is falling during the dry summer months. As a result, variance on the soil moisture is strongly decreasing which is also to be expected. The leaf area index ensemble shows a divergence in the ensemble only after day 100, as a result of soil moisture levels becoming growth limiting. The results for North France demonstrate that the precipitation ensembles can sustain a divergence in the soil moisture results throughout the entire growing season, while the leaf area index results do not diverge up to day 120.

5 Discussion and conclusions

The main result of this paper is a histogram-based approach for transforming heterogeneously distributed precipitation residuals into a Gaussian random variable. The thus transformed
precipitation residuals appeared to be approximately multivariate Gaussian. Similar to results obtained by Kyriakidis et al. (2004), they exhibited strong spatial correlation, while temporal correlation was very weak.

The possibility of having precipitation at locations where CGMS predicted a dry day was handled by an independent indicator simulation which generates rain storms in areas where no precipitation was recorded. The precipitation amount is then obtained through the Gaussian simulation. Currently, this indicator simulation is assumed to be temporally uncorrelated and stationary in space and time. Given the fact that we are interested in precipitation residuals rather than the events themselves, this seems a reasonable approach. In case temporal correlation turns out to be important then the indicator simulation could be conditioned on the simulation results of the previous day.

Currently, our approach assumes stationary processes in space and time for both the Gaussian simulation and the indicator simulation. Several observed deviations from the target properties (ELDAS) may be attributed to this design choice. Our approach reproduced the histogram of the entire dataset, but it does not reproduce the histogram or other statistical properties of any particular location or time-instant. An example is that the approach does not reproduce dry-spell lengths in Spain because too many precipitation events are generated during summer. Although this is an unfavourable characteristic of the proposed simulation approach, its effect should not be overestimated. The results in figure 4 are based on binary sequences of rain/no-rain events which do not take into account the amount of precipitation. In practice, the generated precipitation events dividing long dry periods mainly concern small precipitation amounts.
Preliminary results from the crop model simulations demonstrate that the precipitation ensembles can be used to generate realistic soil moisture ensembles. Obviously, divergence in the leaf area index results is only guaranteed when the soil moisture is limiting for crop growth.

Acknowledgements
The authors acknowledge the helpful comments of Paul Torfs during the development of the methodology and during the finalization of this manuscript. Franz Rubel is acknowledged for providing the ELDAS precipitation database.

References


