Representing Uncertainty in Continental-Scale Gridded Precipitation Fields for Agrometeorological Modeling

A. J. W. De Wit, S. De Bruin, and P. J. J. F. Torfs

Wageningen University and Research Centre, Wageningen, Netherlands

(Manuscript received 2 March 2007, in final form 30 November 2007)

ABSTRACT

This work proposes a relatively simple methodology for creating ensembles of precipitation inputs that are consistent with the spatial and temporal scale necessary for regional crop modeling. A high-quality reference precipitation dataset [the European Land Data Assimilation System (ELDAS)] was used as a basis to define the uncertainty in an operational precipitation database [the Crop Growth Monitoring System (CGMS)]. The distributions of precipitation residuals (CGMS/ELDAS) were determined for classes of CGMS precipitation and transformed to a Gaussian distribution using normal score transformations. In cases of zero CGMS precipitation, the occurrence of rainfall was controlled by an indicator variable. The resulting normal-score-transformed precipitation residuals appeared to be approximately multivariate Gaussian and exhibited strong spatial correlation; however, temporal correlation was very weak. An ensemble of 100 precipitation realizations was created based on back-transformed spatially correlated Gaussian residuals and indicator realizations. Quantile-quantile plots of 100 realizations against the ELDAS reference data for selected sites revealed similar distributions (except for the 100th percentile, owing to some large residuals in the realizations). The semivariograms of realizations for sampled days showed considerable variability in the overall variance; the range of the spatial correlation was similar to that of the ELDAS reference dataset. The intermittency characteristics of wet and dry periods were reproduced well for most of the selected sites, but the method failed to reproduce the dry period statistics in semiarid areas (e.g., southern Spain). Finally, a case study demonstrates how rainfall ensembles can be used in operational crop modeling and crop yield forecasting.

1. Introduction

Process-based mechanistic crop models are an important tool for assessing the effects of crop management, weather, and soil on crop growth. Although crop models were often originally designed and tested at the plot scale, they are nowadays applied in systems with typical spatial resolutions of 0.5–2.5°, and their aggregated output is used to predict crop yield and production at regional, national, and continental scales. Information on the outlook on yield and production of crops over large regions is essential for government services dealing with import and export of food crops, for agencies with a role in food relief, for international organizations with a mandate to monitor world food production and trade, and for commodity traders.

Given the scales at which these systems operate, the model simulation results are subject to large uncertainties and weather input is generally acknowledged to be the largest source of uncertainty (Aggarwal 1995; Easterling et al. 1998; Mathe-Gaspar et al. 2005; Mearns et al. 2001). The main reason for this is that regional crop yield simulators are typically not used to predict crop yield directly; rather, they represent the annual responses of crops to weather variability. These are translated into crop yield forecasts using regression models that relate historic time series of crop yield statistics to simulated crop yields. Consequently, the influence of relatively stable factors like crop variety, soil, and management is largely captured by the trend component in the regression model, and the year-to-year variability in predicted crop yields originates from the simulation model and is mainly due to the weather.

Current operational yield forecasting systems are generally deterministic in nature and not capable of quantifying uncertainties that are inherent in all parts of the system. A shift is therefore necessary toward
probabilistic systems, which are commonplace nowadays in meteorological and hydrological applications. Of particular interest is a comparison with hydrological land surface models, which, similar to crop models, simulate processes such as evapotranspiration, plant growth, and soil moisture dynamics. Two important insights can be learned from probabilistic applications of land surface models: (i) rainfall is the key input for modeling the hydrologic state of the land surface (Syed et al. 2004) and (ii) uncertainty of an available rainfall product is commonly represented by an ensemble that is next used for error propagation assessment (Carpenter and Georgakakos 2004; Crow 2003; Reichle et al. 2002; Seo et al. 2000). A first step for ensemble-based probabilistic crop modeling would be to develop a rainfall ensemble tailored to the temporal and spatial scale required for regional crop modeling.

Although land surface models often operate at time steps of an hour or less, crop growth models typically employ a much coarser time step of one day or more. This is appropriate because, unlike land surface models, their focus is on the cumulative effect of rainfall events on soil moisture content of the root zone (Robertson et al. 2007). For example, the Food and Agriculture Organization (FAO) crop water satisfaction index is a simple crop model that basically relates seasonal accumulated evapotranspiration to accumulated precipitation (Frère and Popov 1986; Verdin and Klaver 2002).

Choosing an appropriate spatial scale is important because of the nonlinear behavior of crop models to weather inputs and the resulting errors that may occur when aggregating model output to administrative regions (Hansen and Jones 2000). Existing studies are not entirely consistent on this aspect. Easterling et al. (1998) report maximum correlation between simulated and observed yield for maize in the U.S. Great Plains when weather data at 100 × 100 km² resolution were used as input. Challinor et al. (2003) found maximum correlation between rainfall data and ground nut yield over India at a scale roughly corresponding to 250 × 250 km², and de Wit et al. (2005) found linear scaling of crop model simulated biomass when precipitation and radiation inputs were scaled from 10 × 10 km² to 50 × 50 km². In contrast, Oleson et al. (2000) found little influence of the scale of precipitation inputs on the explanatory power of winter wheat predictions in Denmark, but this was attributed to the dominating effects of diseases, pests, and harvest conditions. The above-mentioned results are not conclusive on the spatial scale, but they suggest that a resolution of 50 × 50 km² to 100 × 100 km² is relevant for regional crop modeling.

The many rainfall ensemble generation approaches found in the literature can be divided into two main categories. In the first category, ensemble techniques are used to generate synthetic time series of precipitation to represent the natural variability in the precipitation process (Wilks 1999; Yang et al. 2005). In the second category, ensemble techniques are used to characterize the spatiotemporal properties of given rainfall sequences or fields, which are uncertain as a result of limitations in the observed or forecasted data. For example, the ensemble can quantify the uncertainty as a result of limited rain gauge network density or inaccurate precipitation estimates from numerical weather prediction (NWP) models, radar satellites, or ground radar (Bates et al. 1998; Carpenter and Georgakakos 2004; Charles et al. 2004; Clark et al. 2004; Hossain and Anagnostou 2006; Mackay et al. 2001). Similarly, ensemble techniques are used to downscale a coarse-resolution precipitation field to equiprobable precipitation fields at a higher spatial and/or temporal resolution (Margulis and Entekhabi 2001; Seo et al. 2000).

The current paper concerns the second category of rainfall ensemble generators. We propose a relatively simple method for generating ensembles of gridded precipitation residuals at a temporal and spatial scale (daily values and 50 × 50 km² averages) that is consistent with the requirements for crop model applications that target large-area crop yield prediction. The developed methodology does not simulate the precipitation field directly; rather, it computes residual error fields that are added to the input precipitation field to obtain the ensemble trace. This eliminates the need to account for temporal autocorrelation and seasonality of the precipitation fields if it is assumed that the input represents the basic patterns of seasonal, regional, and day-to-day variability. Additionally, the use of residuals avoids the problem of climate variability at decadal scales, given that the decadal variability should be reflected in the measured data rather than in the residuals. This is particularly important for agrometeorological applications, given that often fairly long time series of data need to be generated (>10 yr). The developed method takes into account spatial correlations in the ensemble; it should reproduce the input statistics (mean and variance) and present a practical solution that can be parameterized relatively easily.

For the implementation, we used the weather database of the European Crop Growth Monitoring System (CGMS) as the operational (but uncertain) precipitation product, and we calibrated the ensemble generator using a highly accurate rainfall product available for a limited period. Next, we generated an ensemble of precipitation inputs that characterizes the uncertainty in
the CGMS precipitation product and validated the statistical properties of the ensemble with the reference dataset. We illustrate the use of precipitation ensembles in crop yield forecasting by running the precipitation ensemble through a distributed crop growth model for a district in southern France.

2. Data

a. CGMS meteorological database

An important component of the Crop Growth Monitoring System is the CGMS meteorological database. This database contains daily weather data measured at stations starting in the 1970s and it 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 stations are discontinued or new ones established.

CGMS operates at grid cells of 50 km × 50 km; therefore, a spatial interpolation routine is applied to estimate weather variables for each 50 km × 50 km grid cell (van der Voet et al. 1994). Each cell receives values for temperature, radiation, vapor pressure, evapotranspiration, and wind speed using inverse distance weighting, and rainfall is assigned from the nearest most similar station in terms of elevation and distance to the coast. This method was chosen to avoid the misrepresentation of precipitation sequences caused by averaging values from multiple weather stations. In spite of its simplicity, the CGMS interpolation scheme is known to perform well in terms of accuracy and robustness in comparison with more advanced interpolation schemes (Gozzini and Paniagua 2000). Uncertainty in the CGMS precipitation fields is thus a combined uncertainty as a result of limited station density and rain gauge sampling error, which cannot be separated but which is known to influence the crop model simulation results (de Wit et al. 2005).

b. ELDAS precipitation data

The European Land Data Assimilation System (ELDAS) precipitation database consists of daily precipitation values on a 0.2° grid (~15 km) over Europe for the period 1 October 1999–31 December 2000 (Rubel and 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-time activity and no update is to be expected in the near future. Validation demonstrated that systematic measurement errors for over 90% of the number of stations are within 1 mm day⁻¹. It is important to notice that the ELDAS precipitation estimates are much smoother than a real precipitation field and should be regarded as spatial averages. Nevertheless, given the sheer volume of rain gauge measurements that were used to generate this database, it is considered to provide much better estimates of average daily rainfall than the CGMS meteorological database.

We used the ELDAS database as a reference for modeling the error structure in the CGMS precipitation fields. The ELDAS precipitation database was converted to the 50 km × 50 km CGMS grid by taking the average precipitation of ELDAS cells within a CGMS grid cell (on average, there are 7.5 ELDAS cells per CGMS cell).

c. Exploratory analyses of ELDAS and CGMS precipitation databases

1) Spatial distribution of RMSE between ELDAS and CGMS precipitation databases

We calculated the root-mean-square error (RMSE) between the ELDAS and CGMS daily precipitation values at the resolution of the 0.2° ELDAS grid over the entire period 1 October 1999–31 December 2000. The results (Fig. 1) demonstrate that the spatial patterns in the RMSE are dominated by areas of high RMSE values, which correspond mainly to mountainous areas with west wind-driven precipitation patterns. The precipitation values in the CGMS database do not properly represent the strong temporal and spatial variability of precipitation in these areas, thus leading to higher RMSE values.

We decided to remove grids with high RMSE due to mountainous terrain because these grids are not relevant for agricultural production. CGMS grids were removed using the criterion that the average slope was larger than 3.5° over the 50 km × 50 km grid. Slope was derived from the USGS HYDRO1K dataset (http://lpdaac.usgs.gov/gtopo30/hydro/index.asp) as the average value within a 50 km × 50 km CGMS grid cell. Figure 1 demonstrates that mainly grids in the Alps, Scandinavia, Spanish and Italian mountain ranges, Greece, Turkey, and Romania were removed from the analysis. Although the number of grids that were removed is considerable, there are no important agricultural areas with annual crops located within these grids.
2) Scatterplot of CGMS precipitation versus precipitation residuals

Figure 2 shows the scatterplot of CGMS daily precipitation versus the precipitation residuals (ELDAS minus CGMS) over the period 1 October 1999–31 December 2000. Some important observations can be made from this figure:

- The vertical banding shows that most CGMS precipitation values are integers, whereas the ELDAS precipitation values are real numbers. Based on this observation, we decided to convert all CGMS precipitation values to integer values during further analyses by rounding them to the nearest integer.
- The distributions of the residuals are asymmetric and vary with CGMS precipitation. With low CGMS precipitation, the residuals are mainly clustered near zero, but this tendency vanishes with increasing CGMS precipitation values.

3) Comparison of precipitation distribution classes

An overview of the errors in the CGMS precipitation database is provided by an error matrix between precipitation distribution classes in the CGMS and ELDAS databases (Table 1). The average ELDAS precipitation per CGMS grid was used over the period 1 January 2000–31 December 2000. The number of counts in the first column corresponds to roughly 60% of the total grid–day combinations. To avoid small fractions, all values are given as a percentage of the column total. Precipitation values larger than 80 mm were removed from the analyses because these events are rare.
and not of interest for the application at hand and they would thus needlessly complicate further analysis.

Table 1 lists a large percentage of events during which CGMS precipitation equals 0 mm and ELDAS precipitation is between 0 and 1 mm (47.3%). This may be caused by (i) greater accuracy of the precipitation values in the ELDAS database, (ii) the interpolation used for creating the ELDAS database (block kriging) causing a smoothing of precipitation values, or (iii) the averaging of multiple ELDAS grid points within one CGMS grid.

To compensate for this effect, we decided to treat ELDAS precipitation values lower than 1 mm as 0 mm for events where CGMS precipitation was zero. This procedure effectively adds an additional 47.3% to the 39.6% in the upper left cell of the matrix (Table 1). This decision is justified because the precipitation values lower than 1 mm are insignificant from an agricultural point of view and typically account for only 3% of the yearly total precipitation. A second important observation is that the distribution of ELDAS precipitation within a CGMS precipitation class is non-Gaussian and the standard deviation of the distribution becomes larger with increasing CGMS precipitation.

3. Method

a. Overview of the methodology

The simulation method we developed is based on the simulation of fields of precipitation residuals that can be added to the CGMS precipitation to obtain the ensemble trace (section 3b). The procedure for generating a single daily precipitation realization can be described using three steps:

(i) A Gaussian spatial field is generated using the sgsim program from the Geostatistical Software Library (GSLIB) 2 (Deutsch and Journel 1998), which reproduces the structure of the normal score transformed precipitation residuals [see sections 3c(1) and 3d(1)].

(ii) This Gaussian field is back-transformed into the observed residual space by using the inverse of the normal score transformations that were derived for specific CGMS precipitation intervals [section 3c(1)].

(iii) The spatial field of residuals is then added to the daily CGMS precipitation field to obtain the realization. If the CGMS precipitation value is zero, an indicator field (modeled with the sisim program) is used to modify the residual [sections 3c(2) and 3d(2)]. The procedure and its parameterization are described in more detail in the following sections.
b. Conceptual modeling

We consider Prec(x, t) to be the (unknown) true average precipitation at grid cell x and day t, Prec_{CGMS}(x, t) the precipitation as recorded in the CGMS system, and $\varepsilon(x, t)$ a random spatially and/or temporally correlated residual.

Our spatiotemporal error model for the data concerned is then given by

$$\varepsilon(x, t) = \text{Prec}(x, t) - \text{Prec}_{CGMS}(x, t).$$

(1)

Note that no statement is made about $\varepsilon(x, t)$ having zero mean because $\text{Prec}_{CGMS}(\cdot)$ might be biased [see section 2c(2)].

We assume that a sample or observed realization of the residual variable can be obtained using the ELDAS precipitation data resampled to the CGMS grid $\text{Prec}_{ELDAS}(\cdot)$ as a reference; thus,

$$\varepsilon_O(x, t) = \text{Prec}_{ELDAS}(x, t) - \text{Prec}_{CGMS}(x, t),$$

(2)

where $\varepsilon_O(x, t)$ denotes an observed residual at location x and date t.

Figure 2 is a scatterplot of the observed residuals, which obviously do not follow a Gaussian distribution. 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, that is, $f^{-1}(\cdot)$. Note that by design the transformation functions should handle any bias in the CGMS precipitation data. We thus obtain the following model for the random residual precipitation field $\varepsilon_M(\cdot)$:

$$\varepsilon_M(x, t) = f^{-1}[\varepsilon_O(x, t), \text{args}],$$

(3)

where $\delta(\cdot)$ denotes a correlated standard Gaussian random field and args are any required additional arguments. Section 3c (below) explains that the residuals in cases of zero CGMS precipitation require additional modeling because of the large number of zeros (see Table 1) in the residuals.

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 a standard Gaussian simulation algorithm. By summing the simulated residuals to observed CGMS precipitation data, the required ensemble traces are obtained.

c. Transformation of precipitation residuals

1) Normal score transformation

We tried several mathematical expressions to transform the observed $\varepsilon_O(x, t)$ to a standard Gaussian distribution as a function of $\text{Prec}_{CGMS}(x, t)$, but owing to the irregularity of our dataset we could not find any useful parametric function. Also, checks for seasonal and spatial trends (the latter as a function of the density of weather stations) in the means of the precipitation residuals were unsuccessful with our time series.

We thus decided to use quantile-based normal score transforms for a series of $\text{Prec}_{CGMS}(\cdot)$ intervals. The dataset was divided into 13 $\text{Prec}_{CGMS}(\cdot)$ intervals (listed in Table 1), and for each of these a histogram of the observed residuals $\varepsilon_O(\cdot)$ was produced over the period 1 January–31 December 2000. These histograms were visually compared with histograms constructed over the different seasons as a check on our assumption of absence of a temporal trend. 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 nescore (Deutsch and Journel 1998). Note that the random despiking algorithm that is part of the original program was disabled because it would cause artifacts that would interfere with subsequent analysis (particularly variogram determination).

2) Modeling of zero precipitation

The first bin $\text{Prec}_{CGMS}(x, t) = 0$ required additional processing because the transformation algorithm could not properly transform the large number of zeros [i.e., $\varepsilon_O(x, t) < 1$ mm] in the residuals. Therefore, we introduced a multiplicative (spatially correlated) indicator variable $i(x, t)$ to treat the $\varepsilon_O(x, t) < 1$ mm data, given $\text{Prec}_{CGMS}(x, t) = 0$ mm separately. Hence, in the first bin only the residuals $\varepsilon_O(x, t) \geq 1$ mm are handled by a normal score transform and Eq. (3) was modified to Eq. (4):

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

(4)
In the current work, dependence of \( i(x, t) \) on \( \text{Prec}_{\text{CGMS}}(x, t) \) > 0 and \( \delta(\cdot) \) is not considered.

d. Variogram modeling

1) **NORMAL SCORE TRANSFORMED PRECIPITATION RESIDUALS**

Normal score transformed residuals \( e_O'(x, t) \) were computed for all \( \text{Prec}_{\text{CGMS}}(x, t) > 0 \) mm and for \( \text{Prec}_{\text{CGMS}}(x, t) = 0 \) mm with \( e_O(x, t) \geq 1 \) mm. The data thus obtained were 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 \( \delta(\cdot) \).

2) **ZERO PRECIPITATION INDICATOR VARIABLE**

The zero precipitation indicator variable \( i(\cdot) \) [see Eq. (4)] was considered to be spatially autocorrelated. The indicator variogram was obtained by transforming the precipitation data according to Eq. (5); thus,

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

(5)

The upper option just states that the indicator variable is not used when \( \text{Prec}_{\text{CGMS}}(x, t) > 0 \) mm.

3) **CHECKING TWO-POINT NORMALITY**

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

We used the GSLIB program bigaus to derive the model-induced indicator variograms from the variogram of \( e_O'(x, t) \) [see section 3a(1)]. The experimental indicator variograms were obtained by applying thresholds on the \( e_O'(x, t) \) data as follows:

\[
j(x, t) = \begin{cases} 
1 & \text{if } e_O'(x, t) \leq z(p) \\
0 & \text{otherwise}
\end{cases}
\]

where \( j(x, t) \) denotes an indicator transformed data point and \( z(p) \) is the \( z \) score of the standard normal distribution for quantile \( p \) \((p = 0.10, 0.25, 0.5, 0.75, 0.90)\). Subsequently, date-averaged variogram values were computed from \( j(x, t) \) data that were sampled exhaustively in the spatial domain and at a rate of three per month in the temporal domain.

e. Simulation of residual fields

Figure 3 shows how we implemented Eq. (4) in our simulation method, which needs the CGMS precipitation data, the Gaussian random fields, the indicator random fields, and a set of transformation tables as input. If \( \text{Prec}_{\text{CGMS}}(x, t) > 0 \) mm, then the right-hand branch of the flow diagram suffices (i.e., an unconditional standard Gaussian simulation followed by a back
transform). Otherwise, an unconditional indicator simulation is used to model the event of a positive residual $e_{xt}(t)$, given $\text{Prec}_{\text{CGMS}}(x, t) = 0$ mm. Note, however, that both branches are always executed and that the conditions $\text{Prec}_{\text{CGMS}}(x, t) > 0$ mm or $\text{Prec}_{\text{CGMS}}(x, t) = 0$ mm are handled by postprocessing (bottom of Fig. 3).

The Gaussian and indicator simulations were performed using the public domain sequential simulation programs *sgsim* and *sisim* included in GSLIB 2 (Deutsch and Journel 1998). Both programs employ sequential stochastic simulation, which implies that in random order they simulate the nodes of a grid. Previously simulated nodes are used as conditioning data for subsequent simulations within the same realization if they are within a given search neighborhood.

The back-transforms and postprocessing were performed by the LINT2 module in the TTUTIL library (Kraalingen and Rappoldt 2000) and the Python scripting language. A set of 100 alternative realizations of daily precipitation was generated by adding back-transformed simulated residuals to the CGMS precipitation data.

**f. Evaluation of precipitation realizations**

We evaluated the realizations of the precipitation fields on four different aspects. First, the reproduction of the histograms of the ELDAS precipitation was checked by quantile–quantile (Q-Q) plots of ELDAS and CGMS precipitation against 100 precipitation realizations for selected grids. Second, variogram reproduction of the ELDAS precipitation fields was evaluated for two selected days and five realizations. Third, the rainfall temporal intermittency characteristics for both dry and wet periods were compared for six representative sites for the CGMS precipitation, ELDAS precipitation, and 25 realizations. Finally, we evaluated the grid average cumulative precipitation of the CGMS dataset, the ELDAS dataset, and the rainfall ensembles. Moreover, an error matrix similar to Table 1 was generated to show the distribution of CGMS precipitation classes versus the precipitation realizations over the whole grid and the complete time series.

**g. Probabilistic crop yield forecasting**

To illustrate the use of rainfall realizations in agrometeorological applications, we used the World Food Studies (WOFOST) crop growth model (van Diepen et al. 1989) implemented in the framework of the Crop Growth Monitoring System (Genovese 1998; Vossen and Rijks 1995) to generate an ensemble of yield forecasts for grain maize in the Centre-Est region in southeastern France for the year 2000. First, individual rainfall realizations were used as input for WOFOST simulations to obtain an ensemble of simulated biomass values. This procedure was repeated for all $50 \times 50$ km$^2$ grids in the Centre-Est region, and the resulting time series of simulated biomass values for individual grids were spatially aggregated. The final result of this procedure was an ensemble of space-averaged simulated biomass values that was representative for the region.

In a second step, we used the deterministic version of CGMS for the same crop and region to simulate grain maize biomass values over the period 1992–99. The time series of official reported grain maize yields (EUROSTAT 2005) for this region were used as dependent variables in a regression model with the time trend and the simulated biomass results as independent variables (Supit 1997). The coefficients of this regression model were determined for each dekadal$^1$ time step during the growing season. Finally, the regression models explaining the relationship between reported yield and simulation results over the years 1992–99 were applied in prognostic mode to make a forecast using both the deterministic output for 2000 and the output from all ensemble members for the year 2000.

**4. Results**

**a. Distribution of precipitation residuals for selected CGMS precipitation classes**

Figure 4 shows histograms of the precipitation residuals (CGMS minus ELDAS) for selected CGMS precipitation classes over the whole year (labeled "all") and for the different seasons [December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON)]. Focusing on the histograms of the residuals over the whole year, it shows that at CGMS precipitation zero, the precipitation residuals were zero for about 50% of the data, and a relatively large percentage of precipitation residuals had values between zero and one (Fig. 4a). At a CGMS precipitation of 4 mm, the width of the whole histogram becomes wider, thereby demonstrating larger errors in the CGMS precipitation values (Fig. 4b). At CGMS precipitation values between 10 and 12 mm and between 15 and 30 mm (Figs. 4c,d), the shape of the histogram resembles a truncated normal distribution that becomes progressively wider. This confirms that the magnitude of the errors is related to the precipitation amount.

The histograms for the different seasons demonstrate

---

$^1$ The use of the term dekad is related to an FAO convention to distinguish 10-day periods (dekads) from 10-yr periods (decades).
that there is some variability between the seasons, but there are no obvious trends for particular seasons. The plots suggest that the residuals for the months in spring (MAM) are negatively biased and the residuals during the winter months (DJF) are positively biased in comparison with the yearly histogram (Figs. 4b–d). However, given the absence of any obvious seasonal trends, we decided to use the yearly histograms of precipitation residuals for the normal score transforms.

b. Variogram modeling

1) TRANSFORMED PRECIPITATION RESIDUALS

Figure 5 shows the variograms of normal score transformed precipitation residuals for $\text{Prec}_{\text{CGMS}}(x, t) > 0$ mm. The black dotted lines correspond to variograms of individual dates and the gray continuous line represents the average variogram over all sampled dates. There is considerable spread among the variograms of the individual dates; however, they all indicate that the transformed residuals are spatially correlated with similar ranges. The average variogram was modeled by the sum of two exponential components (Goovaerts 1997, p. 88), one with a partial sill of 0.802 and a practical range of 315 km and another with 0.198 and 6000 km for the partial sill and practical range, respectively. This variogram model was used in the Gaussian simulations. Likewise, the indicator variogram for the data transformed by Eq. (3) (not shown here) was modeled by two exponential components with partial sills and ranges of 0.0372 and 180 km and 0.0828 and 1500 km, respectively.

The level of temporal autocorrelation in $\delta(\cdot)$ was assessed by computing temporal variograms for six CGMS grid cells that were selected to represent climatic variability over the region (southern Spain,
northern Spain, southern France, northern France, central Germany, and Denmark). The temporal variography shown in Fig. 6 did not point at significant temporal correlation of the transformed residuals. Therefore, our model does not account for such correlation.

2) Two-point normality

Based on graphical comparisons of the experimental and Gaussian model-induced indicator variograms of the normal score data (shown in Fig. 7), we decided to accept the assumption of two-point normality in space for the purpose of our study. Nonetheless, the experimental indicator values for the first decile (\( p = 0.1 \)) deviate substantially from the model-induced variogram. This indicates that observed small transformed residuals were more connected in space than under the Gaussian model. However, for the other quantiles (and the vast majority of the data) the fits of the two models are remarkably good. Also in the temporal domain, the density plot of normal score data on subsequent days (not shown) demonstrated that the normal-score-transformed data are approximately bivariate and normally distributed.

c. Precipitation realizations

The reproduction of the histograms of the ELDAS precipitation for the entire year at grid locations in southern Spain (30032), southern France (43044) and central Germany (59061) is shown in the quantile–quantile plots of Figs. 8a–f. Plotting both the Q-Q plots of ELDAS versus CGMS and ELDAS versus 100 realizations allows comparisons between the original CGMS precipitation data and the realizations of the error model.

For the grid in southern Spain (30032), the distributions of ELDAS and CGMS are nearly identical, showing a clustering of points along the 1:1 line (Fig. 8a). Also in the Q-Q plot of ELDAS versus 100 realizations, most points are clustered near the 1:1 line except for the 100th percentile, which is located near the top of the chart (Fig. 8b). The grid in southern France shows a similar pattern, with a slight underestimation of precipitation values by CGMS up to 15 mm and an overestimation of values larger then 20 mm (Fig. 8c). The underestimation was corrected for in the realizations, but the overestimation could not be corrected and was even somewhat amplified (Fig. 8d). Also in this case, the 100th percentile in the realizations is strongly shifted toward the upper part of the chart. Finally, the grid located in central Germany shows a similar pattern, except for a slight overestimation of precipitation values greater than 10 mm and again a shifting of the 100th percentile to higher precipitation values (Figs. 8e,f).

These results demonstrate that the method reproduced the histograms for the selected sites fairly well, but generated too large precipitation values when compared to the largest values present in the original precipitation sequence. This result is caused by the random character of the procedure, which will inevitably sample the tails of the distribution in some realizations, thereby producing large precipitation values.

The cumulative precipitation over the year 2000, averaged over the entire spatial domain, was 667 mm for the CGMS dataset and 797 mm for the ELDAS dataset, suggesting underestimation of precipitation by the CGMS dataset. However, the average of the realizations (generated on the basis of the CGMS product) was 769 mm, demonstrating that the algorithm is able to
adjust the underestimation of precipitation by the CGMS product considerably.

The variograms of 100 realizations at 11 January, 11 April, 11 July, and 21 November 2000 (dotted thin lines), as well as the average variograms over all realizations (thick black line), are shown in Fig. 9. The figures demonstrate that the ranges are fairly well reproduced but there is considerable variability in the overall variance (the sill) of the realizations, which can be both smaller and larger than the ELDAS variance (dashed thick line). For 11 January (Fig. 9a) and 21 November (Fig. 9d), the average of the realizations matches quite well with the ELDAS variogram; for 11 July there is a small overestimation (Fig. 9c); and there is considerable overestimation on 11 April (Fig. 9b). These results could indicate that the algorithm performs better in reproducing the spatial structure for the autumn and winter months than for the spring and summer months.

At small ranges nearly all realizations had larger semivariance compared to the target variance, indicating larger variability in precipitation realizations at small ranges compared to ELDAS. This is directly related to the smoothness of the ELDAS variogram; for 11 July there is a small overestimation (Fig. 9c); and there is considerable overestimation on 11 April (Fig. 9b). These results could indicate that the algorithm performs better in reproducing the spatial structure for the autumn and winter months than for the spring and summer months.

The intermittency characteristics of the dry periods were determined for six sites located in areas with major agricultural production (Fig. 10). The results for southern 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 is not reproduced in the realizations. This is caused by the indicator realizations, which may enforce precipitation events for zero CGMS precipitation. Note that the indicator simulation was configured on average data rather than on-site and season-specific data. Similar effects (albeit less pronounced) can be observed in the results for northern Spain and southern France (Figs. 10b,c).

The results for the sites located in more temperate climate regions (northern France, central Germany, and Denmark) demonstrated that the simulation approach performed better in reproducing the dry period statistics. For these three sites, there are realizations which reproduce the maximum, or even larger, dry period length, although the dry period lengths in the realizations are, on average, still too short.

The intermittency characteristics of the wet periods were determined for the same six sites (Fig. 11). 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 wet day sequences greater than or equal to 1 day can be observed for nearly all sites.

Figure 12 shows precipitation maps for 11 July 2000 according to the CGMS, ELDAS, and the first of a series of realizations of our model. The map of the precipitation realizations shows that large precipitation amounts were drawn from the error distribution for locations in western Germany and Belgium as well as in the eastern part of the Alps. In northern Finland, small
FIG. 8. Q-Q plots of precipitation quantiles for precipitation time series of three selected grids in (a), (b) southern Spain (30032), (c), (d) southern France (43044), and (e), (f) central Germany (59051): (a), (c), (e) ELDAS precipitation sequence vs CGMS precipitation sequence and (b), (d), (f) ELDAS precipitation sequence vs 100 realized precipitation sequences.
rainfall amounts were generated in locations where no precipitation is present in the CGMS gridded precipitation, demonstrating the effect of the zero precipitation modeling branch of the algorithm, which creates occurrences of precipitation in dry areas.

The distribution of CGMS precipitation classes versus the precipitation realizations over the whole grid and the year 2000 are averaged over 100 realizations (Table 2). A comparison of Tables 1 and 2 demonstrates that our method reproduces the target distributions (ELDAS) well. The only exception is that for CGMS precipitation classes larger than 15 mm there was an increase in the number of zero precipitation occurrences in the ensembles. This is caused by the large width of the precipitation intervals used in the back transform, which may produce negative precipitation values. In our current implementation, negative precipitation amounts are set to zero.

d. Probabilistic crop yield forecasting

The regression models that were established between historic EUROSTAT reported yields and the simulated biomass values could explain a considerable percentage of variance in the yield statistics, with $R^2$ values starting at 0.57 in dekad 16, then gradually increasing up to 0.83 in dekad 26 and finally slightly decreasing to 0.78 in dekad 30. The regression was significant starting at dekad 21 (significance level $\alpha = 0.05$) up to dekad 30.

The probabilistic yield forecasts (Fig. 13) based on the regression models demonstrate that the uncertainty in precipitation has a profound influence on the value of the yield forecast during the growing season. Compared to the deterministic yield forecast (blue line), the probabilistic yield forecasts shows a diverging ensemble of yield forecasts, which keeps diverging almost up to end of growing season with maximum spread in the yield forecast of around 0.65 ton ha$^{-1}$. Given that EUROSTAT uses a tolerance of 0.2 ton ha$^{-1}$ as an acceptable accuracy for yield forecasts (G. Genovese 2006, personal communication), this is a significant deviation. The average ensemble yield forecast is clearly higher than the deterministic forecast. However, this is probably due to underestimation of precipitation in the CGMS dataset (de Wit et al. 2005). This bias is corrected in the precipitation ensemble but not in the forecast regression. Finally, the density plots

![Fig. 9. Variograms of the ELDAS precipitation data (dashed thick line), the simulated precipitation fields (dashed lines; 100 realizations) and the average variogram over all realizations (solid thick line) for four selected dates over the year: (a) 11 Jan 2000, (b) 11 Apr 2000, (c) 11 Jul 2000, and (d) 21 Nov 2000.](image-url)
demonstrate that the shape of the yield forecast ensemble is non-Gaussian and propagates in a nonlinear way.

It should be noted that the current analysis does not include a residual error for the forecast regression; it only includes the error as a result of error in rainfall. Although the former usually decreases with time (the closer to the harvest, the less uncertainty), the latter increases with time because errors accumulate. The ensemble of yield forecasts as shown in Fig. 13 can therefore be considered as a conservative estimate of the uncertainty.

5. Discussion

This paper presents a methodology for generating precipitation ensembles tailored to the temporal and spatial scale required for regional crop modeling. Given that crop models are relatively insensitive to intermittency of precipitation (Robertson et al. 2007), we used
an additive approach to create ensembles of daily precipitation values. For applications where intermittency characteristics at small time scales (hourly values) are of prime importance, our approach is probably not suitable and other methods should be applied that are more tailored to such requirements (Hossain and Anagnostou 2006).

Our method uses 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, but temporal correlation was very weak. The virtually absent temporal correlation in the residuals indicates that the daily CGMS records captured temporal precipitation patterns relatively well in our time series.

The possibility of having precipitation at locations where CGMS predicted a dry day was handled by in-
dependent indicator simulation, which generates rain storms in areas where no precipitation was recorded. The precipitation amount is then obtained through the Gaussian simulation. Currently, the indicator simulation is assumed to be temporally uncorrelated and stationary in space and time.

A potential drawback of our approach is that the Gaussian simulation and the indicator simulation are implemented as independent processes. Therefore, it is not guaranteed to produce small precipitation amounts near dry sites and new precipitation events along the fringes of existing wet areas. Dependency between these processes will be difficult to implement within the current framework, given that the precipitation amounts are not simulated directly but are only retrieved by postprocessing the results of the Gaussian simulation. Yang et al. (2005) also signaled this problem, and they pointed out that currently available simulation methods addressing this problem suffer from other drawbacks because these typically assume the same spatial correlation structure for occurrence and amount of rainfall. The extent to which this problem propagates through applications is not clear yet.

Currently, our approach assumes stationary processes in space and time for both the Gaussian simulation and the indicator simulation which use one vari-
ogram model each and a single set of back-transforms. Consequently, it reproduces the histogram of the entire dataset, but it does not necessarily reproduce the histogram or other statistical properties of any particular location or time instant. Many of the deviations from the target properties (ELDAS) that we found during our evaluation of the precipitation realizations can be related to this design choice.

One example of this 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 unfavorable characteristic of the proposed simulation approach, its effect should not be overestimated. The results in Fig. 10 are based on binary sequences of rain/no-rain events that do not take into account the amount of precipitation. When the wet day threshold is raised to 2.5 mm day\(^{-1}\), the situation greatly improves. Another example is the consistent shifting of the 100th percentile in the distribution of precipitation realizations (Fig. 8) to large precipitation values. Similar effects on monthly total precipitation have been described by Margulis and Entekhabi (2001), who suggested resampling of the realizations to select only those realizations with the desired characteristics.

Finally, an important aspect in ensemble generation is the performance of the system in terms of computation time. Our system is build around parts of the TTUTIL library, the sgsim and sisim tools from the GSLIB library, and the MySQL database. These com-
ponents are glued together using the Python scripting language. Currently it takes 7.5 min to generate 100 realizations for 1 day over the full CGMS grid (11 330 nodes) on a 2.6 GHz PC running GNU/Linux. For operational use, realizations would only need to be generated incrementally and this performance would be sufficient. However, generating a full year of data (100 realizations) for retrospective analyses takes nearly 46 h.

Profiling of the application showed that of the different steps within the simulation, 61% of the processing time was used for generating the Gaussian and indicator simulations with gsism and sisim, 31% was used for postprocessing the Gaussian and indicator simulations using Python, and only 8% was used for database communication and miscellaneous tasks. Performance could thus be greatly increased if more efficient methods, such as a modified turning bands algorithm (Mellor et al. 2000), were used for simulation of random fields.

6. Conclusions

This paper presents a methodology for characterizing and quantifying uncertainty in gridded precipitation fields through an ensemble of precipitation realizations with a specific application as the target: regional agrometeorological modeling for crop yield prediction. We defined the temporal and spatial scales which are relevant for this application and developed a relatively simple histogram-based approach that generates residual error fields that are added to the input precipitation field to obtain the ensemble traces.

We calibrated our method using a highly accurate precipitation database and applied it to the precipitation database of the European Crop Growth Monitoring System. The histograms, intermittency characteristics, and spatial structure of the rainfall field were reproduced reasonably well in the realizations, and the deviations that were found (e.g., shifting of 100th percentile, failure to produce prolonged dry spells) are of minor importance for the application at hand.

Finally, we demonstrated that the uncertainty in input precipitation fields and the resulting variability in crop model simulation results considerably influence the yield forecast for a region in southern France. These results demonstrate that there is considerable potential benefit from a probabilistic approach in agrometeorological modeling and crop yield forecasting. Such an approach could be an important support to quantitative risk analyses in a decision making process.

Acknowledgments. The authors acknowledge the help of Franz Rubel for providing the ELDAS precipitation database.

REFERENCES


Kraalingen, D. V., and C. Rappoldt, 2000: Reference manual of
the FORTRAN utility library TTUTIL v. 4, Wageningen University, 97 pp.


——, K. Brugger, P. Skomorowski, and M. Kottek, 2004: Daily and 3-hourly quantitative precipitation estimation for ELDAS. ECMWF/ELDAS Workshop on Land Data Assimilation, Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 19–32.


